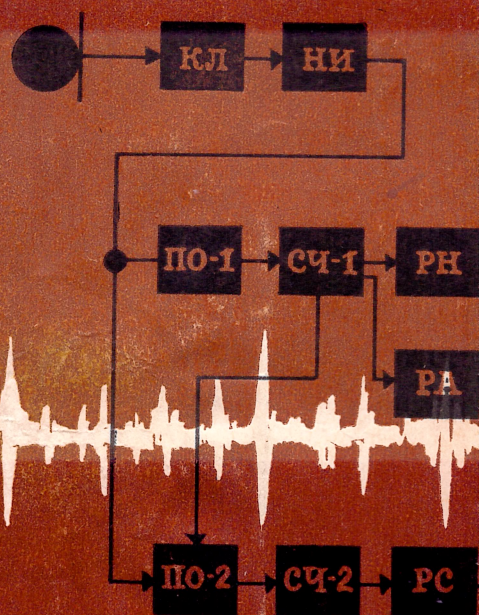


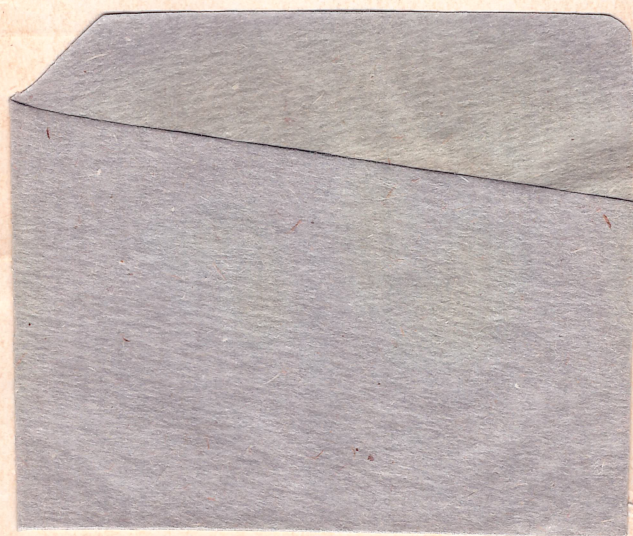
621.391

Ц 36

Г.И. ЦЕМЕЛЬ

ОПОЗНАВАНИЕ РЕЧЕВЫХ СИГНАЛОВ





Академия наук СССР

**Институт проблем передачи
информации**



621.391
Ц 36

ОПОЗНАВАНИЕ РЕЧЕВЫХ СИГНАЛОВ

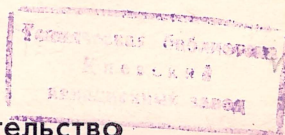
Г. И. Цемель

50198



к

1971



Издательство

«Наука»

Москва

621.391

-2017



Опознавание речевых сигналов. Ц е м е л ь
Г. И., изд-во «Наука», 1971.

Излагаются результаты экспериментальных исследований по опознаванию речи на уровнях сегментов, звуков, слов и фраз с помощью ЦВМ. В основу метода опознавания положены характерные свойства речевого сигнала, вытекающие из способа и места образования звуков речи, и некоторые закономерности восприятия речи человеком. Большинство признаков выделяются с помощью аналогового устройства из клипированного сигнала, в связи с чем исследовалась разборчивость отдельных звуков после клипирования. Разборчивость звуков в этом случае остается достаточно хорошей при пропуске полосы частот до 7 кГц и некотором подавлении нижних частот. Получена следующая надежность опознавания на различных уровнях: для гласных — 88,2% по 1250 реализациям 25 лиц, для щелевых — 91% по 2240 реализациям 46 лиц, для набора 58 слов — 93% по 1060 реализациям 20 лиц; для фраз, содержащих задания на вычисление математических выражений, — 90,2%, с одним переспросом — 98,5%, по 400 реализациям фраз (в среднем с 11,75 слов во фразе) 20 лиц. Излагаются также основные положения и анализируется современное состояние рассматриваемой проблемы. Таблиц 22. Библ. 230 назв. Иллюстраций 64.

Ответственный редактор

доктор техн. наук

И. Ш. Пинскер

ПРЕДИСЛОВИЕ

К настоящему времени создались условия, когда результаты исследований по проблеме автоматического опознавания речи могут наконец выйти из недр лабораторий и найти практическое применение. С одной стороны, в ряде случаев назревает необходимость в установлении речевой связи с управляющими, вычислительными и информационными машинами с тем, чтобы управлять голосом различного рода системами и процессами, получать решения устных заданий (например, на вычисление математических выражений), получать информацию о параметрах систем и различного рода справки. С другой стороны, на базе компактных аналоговых устройств, выделяющих признаки речевого сигнала, и вычислительных машин может быть осуществлено опознавание десятков и сотен слов, а также смысла типовых фраз-заданий, составленных из этих слов. Изложением основных положений рассматриваемой проблемы, путей решения ближайших разрешимых задач и достигнутых к настоящему времени результатов автор надеется привлечь внимание специалистов к вопросам практического использования речевого управления там, где это может оказаться полезным.

Книга начинается с краткого изложения современного состояния проблемы опознавания речевых сигналов. Обзоры по отдельным вопросам этой проблемы даны в соответствующих разделах книги. Литература, относящаяся к различным аспектам проблемы опознавания речи, весьма обширна и к тому же разбросана по многочисленным периодическим и непериодическим изданиям различного профиля. Поэтому обзорные параграфы книги ограничиваются кратким описанием и упоминанием главным образом экспериментальных исследований по проблеме, имеющих техническую направленность, и не претендуют на исчерпывающую полноту. Дальнейшее содержание книги достаточно полно представляет ее оглавление.

Работы, описываемые в книге, начались в конце 1955 г., когда А. А. Харкевич предложил автору заняться исследованием клипированного речевого сигнала с тем, чтобы выяснить возможность использования информации о нулевых пересечениях сигнала для компрессии и опознавания речи. Акад. Харкевич проявлял и в дальнейшем большой интерес к этой проблеме, принимая участие в выборе направления исследований и постановке задач. При об-

суждении результатов первых работ он дал ряд ценных предложений и критических замечаний.

Автор отмечает с благодарностью постоянную поддержку и полезные предложения, которые он имел от Л. А. Варшавского и М. А. Сапожкова.

Большую роль при проведении многих исследований, описанных в книге, сыграл В. Н. Трунин-Донской (Вычислительный центр Академии наук СССР), начавший сотрудничать с автором в 1963 г. Это сотрудничество оказалось весьма плодотворным и привело к широкому использованию вычислительных машин и к расширению фронта исследований, проводившихся ранее автором с помощью одного лаборанта.

Автор также благодарен Г. Я. Высоцкому, А. А. Григоряну, С. Н. Кринову и Б. Н. Рудному за проявленный интерес к проблеме опознавания речи и активное творческое участие в отдельных описанных в книге исследованиях, в том числе в большой трудоемкой работе по построению устройств для выделения признаков речевого сигнала, изготовлению и наладке многочисленных схем, разработке алгоритмов и программ, обработке тысяч реализаций слов и фонем. В этой работе принимали также участие лаборанты А. А. Ерошкин, М. В. Кулагин и работавшие ранее И. И. Колбин, Н. Н. Куликов, В. Н. Нивина и В. Ф. Степанов.

Автор считает своим приятным долгом выразить признательность проф. М. А. Сапожкову и проф. Л. А. Чистович за ряд ценных замечаний по рукописи, способствовавших улучшению книги.

Г. Цемель

ОСНОВНЫЕ ПОЛОЖЕНИЯ ПРОБЛЕМЫ ОПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

§ 1. Современное состояние проблемы

Первая публикация, посвященная автоматическому опознаванию звуков речи, появилась в 1943 г. (Мясников [66]). Действие описанного в этой статье опознающего устройства было основано на распределении энергии речевого сигнала по спектру. Признаками отдельных звуков являлись комбинации знаков разностей энергии в 14 полосах частот, взятых попарно. Надежность опознавания гласных составляла 75—80%.

В последовавших затем исследованиях (Дрейфус-Граф [139, 140], Смит [215], Петерсон и Барни [202], Халле, Хьюз и Редли [162], Варшавский [8], Чистович [105], Фрай и Динес [135, 155] и других) стала заметной тенденция к переходу от попыток опознавания всех фонем к опознаванию некоторых отдельных фонем, главным образом гласных. Определение принадлежности анализируемых звуков, выделенных, как правило, исследователями из ключевых слов (т. е. без учета разнообразия контекста), производилось по характерным спектральным областям или спектральной характеристике в целом. Большую роль в понимании структуры речевого сигнала сыграл спектрально-временной анализ речи с помощью динамических спектрограмм типа «видимая речь» (Поттер, Копп и Грин [205]). В результате проведенных исследований выяснилось, что структура речевого сигнала более сложна, чем это предполагалось ранее. Между фонемами и спектральными характеристиками соответствующих им сигналов не существует простых соотношений. Значительная информация о фонемах содержится в переходных участках сигнала на стыках между звуками. Возникла проблема сегментации непрерывного речевого потока на отрезки, соответствующие фонемам или их частям. Выяснилась необходимость использования лингвистической информации.

Все эти факторы показали, что полное решение рассматриваемой проблемы — опознавание любой связной речи произвольного лица — не может быть достигнуто в ближайшем будущем. Это обстоятельство определило направления дальнейших исследований, проводимых со все возрастающим использованием вычислительных машин. Одним из важнейших направлений является опознавание ограниченного набора слов. Вначале это были машины, опознающие 10 цифр, большей частью с настройкой на голос говорящего (Дэ-

вис, Биддульф и Балашек [132], Дадли и Балашек [143], Форджи [152], Какауридзе и Доценко [47, 48], Динес и Метью [136]). В последние годы словарь опознаваемых слов, произносимых произвольными лицами, расширился до нескольких десятков (Голд [157]; Высоцкий, Рудный, Трунин-Донской, Цемель [18, 19]). Как правило, любое произнесенное слово опознается как одно из слов набора. Но при опознавании с переспросом решение может не приниматься, если максимальное «число очков», указывающее наиболее вероятное слово, ниже порогового значения или если «число очков» следующего слова близко к максимальному. В этих случаях решение не принимается также для некоторых реализаций слов, не входящих в набор.

Наряду с расширением словаря и повышением надежности действия стало заметно стремление к более полному удовлетворению требований практики. Опознающие машины рассчитываются на действие от голосов произвольных лиц (без предварительной настройки на голос говорящего) и на работу в реальном масштабе времени. При соблюдении указанных требований достигнуто опознавание 58 слов с надежностью 93%. Применение простой семантической информации позволяет повысить надежность опознавания этих слов в типовых фразах до 98,5—99,15% [20, 84].

Задача сегментации речевого сигнала решена к настоящему времени для нескольких частных случаев. Достаточно хорошо разделяются сочетания глухой согласный — гласный, несколько хуже — сочетания звонкий шумный — гласный и шумный — сонорный. Большей частью сегментация производится по нескольким признакам, характеризующим способ образования звуков, что позволяет выявить лишь часть границ между сегментами (фонемами) (Уайрен и Стаббс [229], Сакаи и Дошита [208], Цемель [98, 102], Хемдал и Хьюз [169]).

Работы по опознаванию фонем стали характеризоваться использованием в качестве учебной выборки разнообразного речевого материала с тем, чтобы при разработке системы признаков учитывать также влияние контекста. Удовлетворительные результаты (надежность действия порядка 90% при нескольких десятках дикторов и разнообразном речевом контексте) достигнуты при опознавании ударных гласных и глухих щелевых (описаны в гл. 3). С меньшей надежностью опознаются звонкие щелевые и глухие взрывные. Наибольшие трудности возникают при попытках получить признаки сонорных согласных.

Уже при опознавании нескольких десятков слов необходимо частично осуществлять сегментацию речевого сигнала и опознавать некоторые длительные звуки. По мере дальнейшего расширения словаря требуется более полно осуществлять сегментацию сигнала и опознавать все большее количество фонем и их вариантов. Решение задачи опознавания набора слов начинается, таким образом, приближаться к полному решению рассматриваемой проблемы. Однако

при опознавании связной речи (в отличие от сколь угодно большого набора слов) возникает необходимость членения непрерывного речевого потока на слова. Подобное членение до опознавания навряд ли осуществимо. Эксперименты Лисенко [60] показали, что человек не может членить речь на слова по просодической информации в омонимических предложениях (типа: Тамарка упала; там арка упала и другие).

При построении системы опознавания большого набора слов весьма затруднительно сочетать и использовать результаты различных исследований, в частности по опознаванию фона, так как в разных работах используются различные анализирующие устройства, способы сегментации, признаки и алгоритмы принятия решения. Кроме того, отдельные исследования различаются условиями эксперимента (наличием или отсутствием предварительной настройки на голос диктора, степенью учета контекста, временем принятия решения и т. п.).

В ряде случаев возникает необходимость автоматического опознавания (узнавания) личности говорящего. При решении этой задачи приходится опираться на индивидуальные различия голосов разных лиц [49, 74, 75, 111, 164, 188]. Как отмечается в последней статье, различия голосов отдельных лиц не столь категоричны, как, например, отпечатки пальцев. Попытки более точного определения признаков голосов разных лиц затрудняются непостоянством значений этих признаков, зависящих от состояния говорящего. Задача опознавания личности говорящего в этой книге не рассматривается.

Ближайшей разрешимой задачей по данной проблеме можно считать опознавание нескольких сот слов и ограниченного числа типов фраз, составленных из этих слов. До рассмотрения путей, ведущих к решению этой задачи, приведем основные данные о различных факторах, влияющих на характер речевого сигнала, и о некоторых закономерностях восприятия речи человеком, которые должны быть учтены при построении опознающих систем. В настоящем кратком обзоре рассмотрены главным образом работы, имеющие техническую направленность. Другие аспекты проблемы опознавания речи изложены в [5, 24, 35, 59, 78, 88, 94, 109].

§ 2. Влияние характера произношения, индивидуальных особенностей и контекста на речевой сигнал

Речевой сигнал, соответствующий произнесенной фразе, представляет собой непрерывный сигнал с паузами для дыхания и перед некоторыми согласными. Вследствие плавных изменений положений речевых органов (языка, губ, мягкого нёба) при переходе от одного звука к другому в большинстве случаев не удастся точно установить границу между двумя соседними звуками и расчленив таким образом сигнал на последовательность отрезков, характе-

ризующих отдельные звуки. Отрезки сигнала, соответствующие одной фонеме, могут значительно различаться друг от друга ввиду существенного влияния на характер сигнала индивидуальных особенностей и состояния говорящего; темпа, громкости и интонации речи; степени ясности и отчетливости произношения и, наконец, контекста.

Рассмотрим подробнее указанные факторы, влияющие на характер сигнала. Изменения сигнала вследствие индивидуальных особенностей говорящих вызываются: различиями в геометрических размерах речевого тракта; зависимостью основного тона голоса, положения формантных областей от пола и возраста; наличием диалектов, акцента, особенностей и дефектов произношения. Одни из этих факторов приводят к изменениям параметров звуков, вплоть до появления новых вариантов фонем. В других случаях одни фонемы заменяются другими, что приводит к увеличению разброса параметров слов. Приведем несколько примеров. В произношении многих лиц t' и d' оканчиваются щелевыми элементами, чем приближаются по существу к $ц'$ и $\widehat{дз'}$, формально отсутствующим в русском языке. Шумовая составляющая $в$ в подавляющем большинстве случаев весьма мала и влияет больше на натуральность звучания этого звука, чем на разборчивость. У небольшого числа лиц шумовая составляющая $в$ в сочетаниях с огубленными гласными $у$, $о$ (а в отдельных случаях даже с другими гласными) более значительна и сопоставима по уровню шума с $ф$. У некоторых лиц отмечается шепелявость (в том числе вызванная зубным протезом), при которой $с$ и $з$ звучат похоже на $ш$ и $ж$.

Изменения скорости произношения влияют различным образом на длительность отдельных групп звуков. Меньше всех зависят от темпа речи взрывные, больше всех — гласные. Это обстоятельство затрудняет нормализацию речевого сигнала по длительности на уровне слов. При увеличении громкости произнесения в речевом сигнале обычно начинают возникать линейные и нелинейные искажения. Кроме того, глухие смычки после звонких звуков становятся частично или полностью похожими на тональные как из-за реверберации, так и, возможно, вследствие того, что колебания голосовых связок продолжают некоторое время и после смыкания речевого тракта. При тихой речи возрастает удельный вес акустических шумов помещения со всеми вытекающими отсюда последствиями. Изменения громкости влияют также на частоту основного тона.

Просодические характеристики речи (ударение, интонация) сказываются в увеличении длительности и уровня ударных гласных, в особенности в словах, находящихся под синтагматическим или логическим ударением. Основной тон отдельных отрезков сигнала изменяется в соответствии с характером интонации фразы.

Речевой сигнал зависит в значительной мере также от степени ясности и отчетливости произнесения. Как указывал акад. Щерба

[112], возможно бесконечное количество переходных ступеней в стиле произношения — от абсолютной ясности и четкости до небрежной скороговорки, когда все неударные слоги наполовину «съедаются». В связи с этим различают два типа произношения: полный стиль (когда четко произносятся все слоги) и разговорный стиль. Полный стиль обычно употребляется при выступлении в большой аудитории, при затрудненной слышимости (говорящие разговаривают через перегородку), при обращении к лицам, чем-нибудь занятым или плохо знающим язык, и в ряде других случаев, когда для того, чтобы быть услышанным и понятым, приходится четко произносить все слова. Письменный текст языка базируется на полном стиле. Ниже даны образцы фонетической записи полного и разговорного стиля по Л. В. Щербе [112].

Образец полного стиля в русской транскрипции: «ва-днóй-из-ад-да-лén-ных ў-лиц ма-сквы́, фсé-рам до́-ме збé-лы-ми ка-лón-на-ми а-нтре-соль-ју и-па-кри-вíf-шы-мся-ба-лкó-нам, жы-ла́ нé-ка-гда-ба́-ры-ня вда-ва́, а-кру-жón-на-ја мнó-га-чís-ле-нна́й-двóр-ней».

Тот же отрывок в разговорном стиле, в сильно упрощенной транскрипции выглядит так: вá-днó-йз-ы́-ддá-лénн-ых-ўл-й-цмá-сквы́, фсér-ы́-мдо́-мё збél-ы́-мй-кá-лónн-ы́-мй а́н-трь-соль-ј у́й-пй-крь-вíf-шы́-мсь-ба́-лкóн-ым...» Знак ˘ над гласным указывает на его краткость. На первый взгляд различие между полным и разговорным стилями не такое уж большое. Однако следует иметь в виду, что второй образец дан в сильно упрощенной транскрипции. Фактически большинство фонем при разговорном стиле выражено нечетко и не обладает в полной мере своими характерными особенностями. В ряде случаев имеет место полное пропадание отдельных фонем (например, *четыре* произносится как *чтыре*; *говорит* может превратиться в *грит*). Значителен произвол в произношении неударных гласных.

Влияние контекста — соседних и других близко расположенных звуков — на параметры данной фонемы вызывается следующими особенностями речеобразования. В одних случаях еще до произнесения некоторого звука речевые органы занимают положение, удобное для образования также и следующего за ним звука. В связи с этим параметры предшествующей фонемы, на всем ее протяжении, зависят также и от последующей. Например, согласные перед *у*, *о* приобретают губной характер (лабиализируются), что приводит к смещению энергии сигнала в сторону низких частот. В особенности изменяется характер заднеязычных согласных *г*, *к*, *х*. Варианты фонемы *к* перед задними и передними гласными имеют настолько различные спектральные параметры, что лучше различаются по этим параметрам друг от друга, чем от остальных глухих взрывных [78, 97].

В других случаях влияние соседних звуков больше всего сказывается в прилегающих друг к другу участках, что приводит к образованию переходных сегментов, характеризующих одновре-

менно оба звука. Так, например, начальный переходный сегмент гласных *y*, *o*, расположенных после переднеязычных согласных *d*, *t*, *s*, имеет значительно большее значение второй форманты F_2 по сравнению как с квазистационарным значением F_{2c} этих же гласных, так и с переходными сегментами *y* и *o* после других согласных. Таким образом, параметры переходных (а в меньшей мере и квазистационарных) сегментов гласных зависят не только от самого гласного, но и от места образования соседнего согласного.

Можно привести другие примеры зависимости параметров фонем от контекста. Длительность щелевых перед гласными больше, чем перед глухими взрывными [103]. Гласный перед звонким согласным продолжительнее, чем перед глухим, а перед щелевым продолжительнее, чем перед смычным [65].

Несмотря на перечисленные выше факторы, вызывающие большой разброс акустических параметров фонем и слов, собеседники, как правило, достаточно хорошо воспринимают связную речь, отдельные слова и звуко сочетания (если не считать разговоров по каналам связи, искажающим речевые сигналы). Бывают, конечно, случаи, когда речевое сообщение воспринято неправильно или неполностью, что обычно влечет за собой переспрос. Тем не менее речь остается основной формой непосредственного общения между людьми. Представляется поэтому заманчивым использовать особенности (или даже механизм) восприятия речи человеком при построении устройств для автоматического опознавания звуков речи и слов. Перейдем к рассмотрению некоторых закономерностей восприятия речи и выяснению того, какие из этих закономерностей и в какой мере целесообразно использовать при автоматическом опознавании речи.

§ 3. О некоторых закономерностях восприятия речи человеком

Большую роль при восприятии речи играет понимание смысла сказанного (знание языка). Достаточно услышать лишь часть звуков произнесенной фразы, чтобы понять всю фразу, что является результатом известной избыточности речи. Можно полагать, что это обстоятельство, позволяющее понимать разговорный стиль речи, способствовало и его появлению. Понятия «услышать» и «понять» часто ассоциируются. Если воспринята достаточная для понимания часть слова (или фразы), то говорят, что слово или фраза услышаны. Роль семантической информации при восприятии речи хорошо иллюстрируется увеличением разборчивости слов и фраз по сравнению с разборчивостью слогов при передаче речи по каналам связи. Так, например, при правильном восприятии лишь 60% переданных слогов разборчивость слов составляет ~ 95%, а фраз ~ 98% (Сапожков [78]).

Эти и другие аналогичные данные показывают, что принятие решения при восприятии речи человеком происходит на нескольких уровнях, начиная от фонем и слогов и кончая фразами и сообщением в целом. Неопределенности, двусмысленности и ошибки, имеющие место на низших уровнях, разрешаются в значительной мере благодаря привлечению семантической информации на высших уровнях. Однако эта априорная информация может привести и к ошибочным решениям. Слова, неизвестные слушателю, воспринимаются иногда как известные (близкие по звучанию), если они подходят по смыслу фразы, а редко употребляемые слова — как часто употребляемые.

Некоторые подобные случаи хорошо запоминаются. Однажды в аптеке мне посоветовали: «Вам вполне подойдет *анальгин*». О существовании подобного лекарства я знал. Уплатив указанную сумму, попросил *анальгин*. «Не *анальгин*, а *танальгин*», — услышал я в ответ. Дома рассказываю: «Оказывается, существуют *анальгин* и *танальгин*». Разворачиваю покупку и читаю ... *танальбин*. При таких обстоятельствах мой словарный запас пополнился еще одним словом.

Другой случай. Защита диссертации. Сажу в конце зала. Читывают отзыв Омского политехнического института. Размышляя: кажется, в Омске не было такого института. Чтение отзыва подходит к концу. Слышу знакомую мне фамилию рецензента — проф. Азбукина, читавшего лекции в Москве, а затем в Томске. Сразу становится ясным: отзыв прислан не из *Омска*, а из *Томска*. Оба эти слова встречаются примерно одинаково часто, но энергия звука *т* намного слабее энергии *о*, и ее значения в конце зала оказалось недостаточно для правильного восприятия слова. Не помог в данном случае и начальный переходный сегмент *о*, имеющий после *т* повышенное значение F_2 . Ошибка при восприятии слова была исправлена по содержанию другой фразы сообщения.

Роль семантической информации при восприятии речи на уровне слов, фраз и сообщений в общих чертах ясна. Нас будет больше интересовать (по крайней мере в данное время) акустическое восприятие речи на уровне сегментов, фонем и слогов, отличающееся рядом особенностей по сравнению с восприятием неречевых звуков. Рассмотрим эти особенности.

При восприятии речи человеком учитывается не только абсолютная физическая характеристика каждого данного отрезка сигнала, но и окружающие его участки. Формирование фонемных образов происходит на основе последовательно слагающихся дискретных признаков сегментов (Дукельский [35]). Эта особенность является при прослушивании отдельных сегментов и их сочетаний, выделенных из речевого сигнала. Отрезок, соотносимый с начальным переходным сегментом гласного, звучит нечетко, большей частью в соответствии с его акустической характеристикой, зависящей от места образования соседнего согласного. Но достаточно

несколько удлинить его за счет примыкающей квазистационарной части гласного, чтобы полученный таким образом сигнал воспринимался как близкий соответствующему гласному. Например, в сочетании *та* начальный переходный сегмент гласного длительностью 40—50 мсек звучит как æ (или э), но стоит удлинить его на 20 мсек в сторону гласного, чтобы он воспринимался как *а*. Как пишет Дукельский, при прослушивании подобного удлиненного отрезка звучание переходного сегмента как бы не учитывается. Решающее значение приобретает информация, содержащаяся в конечной меньшей части отрезка. Другой пример — последовательность трех сегментов: глухой смычки, шумового импульса и перехода последующего гласного воспринимается как глухой взрывной звук. Таким образом, механизм восприятия речи не усредняет информацию, заключенную в прослушиваемом отрезке, включающем в себя участки двух и более сегментов, а преобразует ее, имея целью получить фонемный образ.

При изменении естественного порядка следования сегментов на обратный (путем воспроизведения записанных на пленку слов в обратном порядке) восприятие звуков, определяемых по последовательности двух-трех сегментов, также изменяется. Начальные *п*, *т*, *к* в подавляющем большинстве случаев не воспринимаются, а конечные — переходят большей частью в *х* или в соответствующие по месту образования щелевые. Гласные, щелевые и носовые звуки в основном сохраняют при обратном порядке следования свои значения, хотя и теряют в некоторой степени естественный характер звучания.

Как уже указывалось, параметры двух соседних звуков, различных по месту образования, согласовываются переходными сегментами. При нарушении подобной согласованности (соответствующей пересадкой отрезков сигнала после записи его на пленку или при синтезе речи) одни и те же сегменты могут при восприятии образовать различные фонемные образы. Как отмечает Гофман [170], при сочетании в синтезированном сигнале признаков различных фонем часть слушателей воспринимает его как одну фонему, остальные — как другую. Добавим, что в ряде случаев в подобных ситуациях слушателями может восприниматься и третья фонема, признаки которой в сигнале отсутствуют. Для сравнения заметим, что при восприятии зрительного образа, составленного из частей двух различных известных фигур, человек сравнительно легко узнает обе исходные фигуры.

Рассмотрим несколько примеров. При взаимной пересадке различных по месту образования взрывных *т* и *п* в сочетаниях *ту* и *пу* вновь образованные *пу* воспринимались в 70,5% случаев как *ту* и в 29,5% случаев как *пу*, а *ту* — главным образом как *пу* (51,5%) или *ту* (44% оценок) [35]. Эти цифры свидетельствуют о большей значимости переходного сегмента гласного (в сигнале после пересадки) для определения места образования соседнего

взрывного по сравнению с его шумовым сегментом (самим взрывным). Однако при прослушивании в отдельности шумового сегмента и переходов подобных сочетаний роль шумового сегмента более важна [159, 162]. При взаимной пересадке *т* и *к* в сочетаниях *ту* и *ку* вновь образованные *ку* воспринимались в 71,5% случаев как *ку*, в 19,5% случаев как *ту*, а *ту* — как *пу* (56,5%) или *ту* (40%). Несмотря на то, что во втором примере заменен лишь один из четырех «участников» пересадки (*п* на *к*) и что переходные сегменты *у* в сочетаниях *пу* и *ку* близки друг другу, результаты восприятия вновь образованных слогов существенно изменились. Уменьшение доли воспринятых *ту* с 70,5 до 19,5% и замена их на *ку* объясняется активной ролью во втором примере весьма значимого признака — длительности взрывного — *т*, по которому заднеязычные взрывные *к* и *г*, имеющие большее *т*, хорошо различаются от остальных взрывных (см. § 6 гл. 3). Большинство *т* после пересадки не могло восприниматься в соответствии с переходом *у* как *к* (из-за малой длительности *т*) или как *т* (из-за большого различия переходных сегментов *у* в *ту* и *ку*) и поэтому воспринималось как *п*, несмотря на отсутствие *п* в исходных сигналах. Длительные звуки (щелевые, гласные), основная информация о которых содержится в их квазистационарной части, воспринимаются после пересадок в подавляющем большинстве случаев без существенных изменений.

На результаты восприятия речи с пересадкой отдельных сегментов влияет также способ выбора границ сегментов (что нужно учитывать при сравнении результатов разных исследователей). В работе Дукельского [35] членение сигнала основывалось на условии оптимальной опознаваемости каждого из расчленяемых соседних звуков при минимальной опознаваемости каждого второго звука. В другом, более раннем исследовании восприятия взрывных в зависимости от контекста (Шац [213]) взрывные при пересадке выделялись двумя способами: вместе с последующим придыханием и сразу после взрыва. Количественные результаты восприятия приведены только для взаимных пересадок согласного *к* в сочетаниях с *и*, *а*, *і*. В случае пересадки *к* вместе с придыханием слушатели почти единодушно воспринимали его правильно независимо от последующего гласного. Во втором случае основную роль играл гласный (точнее, переходный сегмент гласного); *к* в сочетаниях с другими гласными воспринималось главным образом как *р* и *т*. Результаты восприятия *к* после пересадок, приводимые Дукельским, носят промежуточный характер; роль *к* и последующего гласного примерно одинакова.

Хотя признаки фонем перекрываются и распределены в общем случае на протяжении слога, основной единицей восприятия на акустическом уровне является скорее всего фонема. Опыты по записи отдельно согласных и гласных при прослушивании слогов типа СГ (Чистович, Кожевников, Алякринский и др. [109]) пока-

зали, что человек воспринимает и записывает согласный раньше, чем гласный. В ряде случаев могут восприниматься отдельные сегменты, например тональная смычка звонкого взрывного. По данным Чистович [106] человек начинает имитировать слышимый согласный *б*, *г* или *д* до окончания его произнесения.

Таковы некоторые особенности восприятия речи человеком, которые необходимо учитывать при построении опознающих машин. Рассмотренные особенности касаются результатов восприятия, а не механизма восприятия отрезков естественной или преобразованной речи. Вопрос о том, как слуховая система человека выделяет и измеряет признаки речевого сигнала, обрабатывает полученную информацию и принимает решение на различных уровнях, пока мало изучен. Надо полагать, что и при выяснении в большей степени механизма слухового восприятия человека навряд ли окажется возможным или целесообразным встроить его в ближайшем будущем в машины для автоматического опознавания речи. В то же время в этих машинах могут быть использованы признаки речевого сигнала, не воспринимаемые человеком, в том числе такие, как асимметрия огибающих положительных и отрицательных полуволн сигнала (Дерш [137]), наличие ультразвуковых составляющих в диапазоне до 80 *кГц* (Мясникова [69]). При построении опознающих машин желательно получить возможно большую информацию из речевого сигнала с тем, чтобы в возможно меньшей мере прибегать к использованию лингвистической информации.

§ 4. Выбор и характеристика направления исследований

Скромные (по сравнению с конечной целью) успехи, достигнутые при разрешении проблемы опознавания речевых сигналов, объясняются, как следует из изложенного выше, в первую очередь двумя причинами: 1) сложным характером речевых сигналов, зависящих не только от произносимых звуков, но и от контекста, отчетливости, громкости, темпа и интонации произнесения, а также индивидуальных особенностей и состояния говорящего; 2) трудностью использования необходимых для опознавания речи лингвистических данных — информации о словарном составе, правилах и строе языка, которыми обладает человек, знающий тот или иной язык. Все эти данные в полном объеме настолько обширны и громоздки, что в ближайшее время невозможно вместить их приемлемым образом в опознающее устройство. В связи с этим приобретает важное значение выбор направления исследований.

Проблема автоматического опознавания речевых сигналов является частью более широкого раздела технической кибернетики, рассматривающего вопросы опознавания различного рода образов (слуховых, зрительных и др.). При решении этой проблемы можно основываться преимущественно на позиции общей теории опозна-

вания образов. Имеется значительное число работ, посвященных методам принятия решений по данным классам объектов и характеризующим их признакам. Спектрально-временное описание речевого сигнала содержит почти всю необходимую информацию о речи, и спектральные отсчеты, взятые через некоторые определенные промежутки времени, могут служить признаками сигнала. Однако подобная система признаков оказывается громоздкой и недостаточно эффективной. Весьма информативные и мало информативные признаки анализируются и обрабатываются в одинаковой степени. Можно согласиться с мнением Голда [157], что преобразование «плохих» признаков в «хорошие» с помощью статистических моделей больше подходит для более простых задач (опознавания образов), чем для опознавания речи.

Более целесообразным является подход, учитывающий в первую очередь характерные свойства речевых сигналов (вытекающие из условий речеобразования) и закономерности восприятия речи человеком. Приведем характеристику этого направления исследований. Речевой сигнал рассматривается состоящим из последовательности квазистационарных и переходных сегментов, классифицируемых в соответствии с сочетанием или характером изменения сегментных признаков. Эти признаки должны отражать способ и место образования звука. Исследования речевых сигналов с целью получения сегментных признаков сочетаются в настоящей работе с разработкой использующих эти признаки систем, опознающих наборы слов и типовые фразы (составленные из этих слов). Напомним, что разработка подобных систем является ближайшей разрешимой задачей по данной проблеме, способной найти практическое применение.

В соответствии с многоступенчатой структурой восприятия речи человеком опознавание речи в общем случае должно производиться на уровнях сегментов, фонем, слов и фраз. Решение о фонемной принадлежности отрезка сигнала принимается по нескольким последовательно расположенным сегментам. При опознавании длительных фонем (гласных, щелевых) основную роль играют их квазистационарные сегменты. Слова опознаются по словесным признакам, характеризующим наличие, местоположение, взаимное расположение и соотношения длительностей определенных сегментов и фонем в слове. По мере увеличения числа полученных сегментных признаков оказывается возможным увеличивать количество опознаваемых фонем и слов. Для опознавания небольшого набора слов достаточно применять две системы признаков на уровне сегментов и слов.

Информация о системе фонем и словарном составе данного языка частично используется уже при опознавании ограниченного набора слов. Применение правил построения предложений и некоторой семантической информации становится необходимым при опознавании фраз. Задача внесения в опознающую машину линг-

вистической информации должна решаться постепенно. На первых порах следует применять простые методы использования смысловых связей между словами фразы для исправления некоторых неправильно опознанных слов по смыслу фразы, что повысит надежность опознавания фраз.

При создании опознающих систем возникает весьма важный вопрос о принципах построения аппаратуры для выделения и формирования признаков речевых образов на различных уровнях и для принятия решения о принадлежности этих образов к определенным классам на основании полученных признаков. Было признано целесообразным выделять сегментные признаки компактным аналоговым устройством из отрезков сигнала, соответствующих словам. При получении сегментных признаков широко используются данные, содержащиеся в расположении «нулей» клипированного речевого сигнала. Для формирования признаков на более высоких уровнях и принятия решения на всех уровнях используется ЦВМ. Сегментные признаки, полученные в аналоговом устройстве в цифровой форме, вводятся непосредственно в ЦВМ. Опознавание слов и фраз производится в реальном масштабе времени, на голос произвольного лица (без предварительной настройки на голос говорящего). Слова и фразы должны произноситься полным стилем, фразы, кроме того, — с паузами между словами. Большое внимание уделяется автоматизации процессов выделения признаков и их последующего анализа, так как только таким путем можно обработать большой объем речевого материала, необходимый для получения достоверных данных.

Работа в намеченном направлении началась с исследования признака усредненной плотности нулей ρ_t , позволяющего выделить из непрерывного речевого сигнала длительные шумные звуки. Признак ρ_t был использован для опознавания трех слов: нуль, один, стоп [98]. Добавление к ρ_t других сегментных признаков, разделяющих с некоторым приближением сигнал на отрезки в соответствии со способом образования фонем, позволило опознавать 10 цифр [82, 102]. Проведенные затем исследования щелевых и гласных, завершившиеся получением признаков квазистационарных сегментов этих звуков, дали возможность опознавать несколько десятков слов, а также фразы [18, 20]. Наконец, добавление признаков места образования согласных и наличия смягчения (палатализации) слога, полученных при исследовании переходных сегментов гласных [30], дает основание вести работу по опознаванию нескольких сотен слов.

ИССЛЕДОВАНИЯ КЛИППИРОВАННОГО РЕЧЕВОГО СИГНАЛА

§ 1. Об использовании клиппированного сигнала при опознавании речи

При предельном амплитудном ограничении — клиппировании — речевой сигнал сохраняет лишь два возможных значения. Тем не менее разборчивость клиппированной речи, несмотря на недостаточную натуральность звучания, остается довольно высокой. По данным Ликлайдера [187], разборчивость односложных слов, составлявшая при неискаженном сигнале 99,6%, уменьшилась после клиппирования до 86 и 97,4% (вторая цифра относится к случаю, когда ограничению предшествовало дифференцирование сигнала). Для русской речи разборчивость звуков при степени ограничения 80 дБ уменьшалась соответственно с 97 до 85 и 93% (Ростовцев [76, 77]).

Использование амплитудного ограничения речевых сигналов в целях повышения помехоустойчивости было предложено, по крайней мере, еще в 1937 г. Высоцким и Тетельбаумом [16]. Это предложение основано на увеличении после клиппирования уровня слабых сигналов (соответствующих глухим согласным) до максимальных значений речевого сигнала, имеющих место при гласных звуках. Некоторые применения клиппированных сигналов, главным образом в системах связи с однополосной модуляцией, описаны в [175, 179]. Хорошая разборчивость и двоичная форма клиппированного сигнала делает целесообразным использование его при анализе речевых сигналов с целью выделения признаков для опознавания речи [92, 96, 97, 116, 209, 210]. Образцы осциллограмм отрезков речевого сигнала до и после клиппирования для различных звуко сочетаний даны в Приложении.

В связи с применением в схемах клиппирования значительного усиления различного рода помехи возрастают до уровня полезного сигнала и заполняют паузы шумом, а также искажают сам сигнал. Помехи можно разделить на два вида. К первому из них следует отнести дополнительные звуки, сопутствующие звукам речи, но не характерные для них. В обычной речи эти помехи не замечаются слушателями, а если и заметны, то мало влияют на разборчивость. Ко второму виду можно отнести мешающее действие реверберации и остальные акустические шумы. Помехи второго вида могут быть сведены до минимума установкой микрофона в заглушенной ком-

нате. Однако в подавляющем большинстве случаев опознающее устройство должно работать в нормальных условиях.

Одной из действенных мер по уменьшению влияния помех является ограничение речевого сигнала по минимуму. На выход устройства проходят только сигналы, превышающие некоторый установленный порог ограничения. Все виды помех с меньшим уровнем отсеиваются. Влияние амплитудного ограничения по минимуму на разборчивость звуков речи рассмотрено Быковым [7]. Если допустить потерю разборчивости до 1%, то степень ограничения по минимуму CO_M не должна превышать примерно 0,4 дБ

$$CO_M = 20 \lg \frac{u_{\max}}{u_{\max} - u_{\text{огр}}},$$

где $u_{\max}$ — максимальное напряжение, достигаемое в данной полосе; $u_{\text{огр}}$ — пороговое напряжение, соответствующее началу ограничения.

Задаваясь величиной $CO_M = 0,4 \text{ дБ}$, получим $u_{\text{огр}} = 0,045 u_{\max}$. На практике уровень ограничения может быть также подобран исходя из условий сохранения достаточно хорошей разборчивости и устранения шумов в паузах.

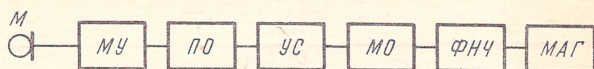
С целью уменьшения мешающего действия реверберации, ослабления низкочастотных акустических шумов, а также улучшения разборчивости согласных нижние частоты в речевом тракте до ограничения должны быть несколько подавлены.

Пониженная разборчивость отдельных фонем не имеет решающего значения при прослушивании человеком связной речи. При автоматическом опознавании речевых сигналов необходимо получить возможно большую информацию на акустическом уровне с тем, чтобы в возможно меньшей степени прибегать к использованию лингвистической информации. В связи с этим желательно выявить признаки всех отдельных фонем. Возникает вопрос, достаточно ли информации об отдельных фонемах остается в сигнале после клиппирования. Заметим, что приведенные в [76, 187, 211] данные характеризуют разборчивость клиппированной речи лишь в целом. Для выяснения этого вопроса были проведены описываемые ниже испытания по определению разборчивости фонем клиппированной речи.

§ 2. Влияние клиппирования речевого сигнала на разборчивость фонем

Для проведения исследований по разборчивости фонем были составлены 40 испытательных таблиц, содержащих по 25 звуко сочетаний типа СГ (согласный — гласный) и ГС. Блок-схема речевого тракта, использованного при этих исследованиях, приведена на рис. 1. Шумы в паузах подавлялись применением ограничения

по минимуму. Порог ограничения соответствовал звуковому давлению $0,003 \text{ н/м}^2$ при чувствительности микрофона $0,25 \text{ мв/0,1 н/м}^2$. Испытания на разборчивость проводились группой в составе 5 человек, которые не проходили какой-либо специальной тренировки.



Р и с. 1.

Блок-схема речевого тракта

М — микрофон; МУ — микрофонный усилитель; ПО — симметричный четырехкаскадный пиковый ограничитель; УС — усилитель; МО — ограничитель по минимуму; ФНЧ — фильтр нижних частот 2,7 или 7 кГц; МАГ — магнитофон МАГ-8М

Разборчивость гласных клипированной речи для мужских голосов (по двум дикторам) при ширине передаваемой полосы частот 2,7 кГц приведена в табл. 1.

Т а б л и ц а 1

Разборчивость гласных клипированной речи для мужских голосов в процентах, ФНЧ 2,7 кГц

Передано	Принято						Пропуск
	а	о	у	э	и	ы	
а	98,7	0,2	0,1	0,2	0,1		0,7
о	1,1	86,7	—	7,5	0,2		4,5
у	0,4	—	90,2	—	4,5	1,6	3,3
э	0,7	2,1	—	92,8	1,0		3,4
и	—	—	0,7	—	88,7	7,3	3,3
ы	0,5		4	—	20,5	70	5

Разборчивость гласных, как это и следовало ожидать, довольно хорошая. Хуже других гласных воспринимается *ы*. По разборчивости в полосе 2,7 кГц гласные клипированной речи для женского голоса (табл. 2) можно разделить на две группы. Фонемы *а*, *о*, *э* сохраняют хорошую разборчивость. Остальные гласные — *у*, *и*, и в особенности *ы*, — в значительной мере переходят друг в друга. Следует отметить, что большая часть неправильно принятых гласных относится к слогам, где не были правильно приняты согласные. Для сравнения в табл. 3 приведены значения разборчивости гласных обычной речи, прошедшей через тот же тракт, исключая нелинейные элементы.

Таблица 2

Разборчивость гласных клипированной речи для женского голоса в процентах, ФНЧ 2,7 кГц

Передано	Принято						Пропуск
	<i>a</i>	<i>o</i>	<i>y</i>	<i>э</i>	<i>и</i>	<i>ы</i>	
<i>a</i>	98,8	—	0,2	0,2	—	—	0,8
<i>o</i>	0,3	96,7	—	1,9	—	—	1,1
<i>y</i>	—	0,3	70,9	—	14,8	11,5	2,5
<i>э</i>	—	0,5	—	95,4	—	—	4,1
<i>и</i>	—	—	35,2	—	54,1	10,2	0,5
<i>ы</i>	—	—	51,4	—	12,1	33,6	2,9

Таблица 3

Разборчивость гласных обычной речи для мужских голосов в процентах, ФНЧ 2,7 кГц

<i>a</i>	<i>o</i>	<i>y</i>	<i>э</i>	<i>и</i>	<i>ы</i>
99,5	97,4	100	98,1	96,2	97,7

При расширении полосы частот клипированной речи до 7 кГц заметно улучшается разборчивость лишь фонемы *и* (с 88,7 до 94,5%). Подавление низких частот до клипирования (частота 300 Гц ослаблялась на 6 дБ относительно 3000 Гц) улучшает разборчивость гласных до величин, указанных в табл. 4.

Таблица 4

Разборчивость гласных клипированной речи для мужских голосов в процентах, ФНЧ 7 кГц с подавлением НЧ

<i>a</i>	<i>o</i>	<i>y</i>	<i>э</i>	<i>и</i>	<i>ы</i>
99,8	97,7	97,8	99,3	99,2	96,5

Примечание. 3,5% *ы* были приняты как *и*.

Улучшение разборчивости гласных *y*, *и*, *ы* для женского голоса в полосе 2,7 кГц может быть получено подавлением низких частот до клипирования. Расширением спектра достигается дальнейшее улучшение различения этих гласных до величин, близких к указанным в табл. 4.

Результаты восприятия начальных согласных клиппированной речи для мужского и женского голосов при ограничении спектра частотой 2,7 кГц приведены в табл. 5. Величины разборчивости для мужского и женского голосов для большинства фонем близки друг к другу. В следующей таблице приведены аналогичные данные для обычной речи, прошедшей через тот же тракт, исключая нелинейные элементы. Табл. 6 составлена по небольшому числу слоговых таблиц, и ее данные следует считать приближенными. Из сравнения табл. 5 и 6 следует, что по разборчивости согласных в полосе 2,7 кГц клиппированная речь значительно уступает обычной речи. Лишь 6 фонем (из 32) были правильно приняты в более чем 75% случаев. В обычной речи эту величину превысили 19 фонем. Неправильно принятые фонемы в большинстве случаев относятся

Таблица 5

Разборчивость начальных согласных клиппированной речи в процентах, ФНЧ 2,7 кГц

Передано	Принято правильно	Наиболее частые замены				Передано	Принято правильно	Наиболее частые замены			
б	20,8	в — 33,8	м — 16,5	— 14,3		б'	18	— 56,5	м — 12	в — 11	
д	8,5	б — 26,8	— 20,5	в — 17		д'	1	— 62,8	в — 19,5	н — 4,3	
г	16	в — 29	— 24	м — 8,5							
п	29	— 46,5	к — 18,5	н — 3,5		п'	58,4	к — 16,3	— 12,9	б — 6,3	
т	10,8	к — 47,5	п — 14,8	— 13		т'	24,5	ч — 36,5	к — 32	п — 6,5	
к	71	— 20	п — 3,5	н — 2							
в	50,5	— 28,5	б — 6,2	л — 5,5		в'	8,8	— 69,1	м — 4,8	н — 4,8	
ж	95,8	— 2,7	л — 1								
з	3,5	в — 36,5	— 26	ж — 10,5		з'	6,5	— 68	р — 17,5	л — 6,5	
ф	4,4	ч — 35,5	ш — 21,3	к — 11,7		ф'	5	ч — 62,8	с — 8,4	п — 5,7	
ш	80	ц — 7,3	с — 5,3	х — 3,5							
с	36	ш — 20	— 9,5	ф — 8		с'	53	ч — 18,7	ц — 16,5	х — 4,3	
х	78,5	ч — 6,5	ш — 4,5	с — 3,5							
м	42	— 27,6	н — 13	л — 8,8		м'	53,2	н — 24,9	— 18,5	р — 2,8	
н	31,5	— 29	м — 25,5	л — 6		н'	32,5	— 31	м — 26,5	л — 2	
л	38,2	— 37	р — 7,5	н — 6		л'	34,2	— 33,6	н — 17,1	р — 2,8	
						ј	81,9				
р	96	л — 2	ц — 1	— 1		р'	92,5	л — 5	— 3	ж — 1	
ц	30	с — 16,3	— 10,2	х — 10,1		ц	57	к — 15	с — 8,5	ц — 8,5	

Примечание. Здесь и далее в графе наиболее частых замен знаком «—» (тире) обозначен пропуск при приеме.

Таблица 6

Разборчивость начальных согласных обычной речи в процентах, ФНЧ 2,7 кГц

Пере- дано	Принято правиль- но	Наиболее частые замены	Пере- дано	Принято правиль- но	Наиболее частые замены
б	94		б'	100	
д	74	г — 17	д'	32	г — 54
е	95				
п	60	— 11	п'	79	
т	71	к — 14	т'	7	к — 57
к	89	п — 7			
в	88	л — 7	в'	78	з — 14
ж	97				
з	43	ж — 57	з'	28	в — 43 — 28
ф	57	с — 29	ф'	24	х — 52 ч — 14
ш	100				
с	57	ф — 21	с'	71	х — 17 ф — 9
х	100				
м	64	н — 36	м'	78	
н	81	м — 19	н'	81	
л	95		л'	100	
			ј	99	
р	100		р'	100	
ц	52	с — 17 ф — 14 х — 6	ц	100	

к той же группе по способу образования, что и переданные. Средняя разборчивость согласных клипированной речи, без учета частоты встречаемости, составила 44,2% (для мужского голоса 43,9%, для женского 44,5%). Следует отметить, что связная клипированная речь в полосе 2,7 кГц довольно разборлива.

Расширение полосы частот речевого сигнала после клипирования до 7 кГц значительно улучшило разборчивость начальных согласных, в среднем с 44,2 до 70,6% (табл. 7). В особенности возросла разборчивость д' — с 1 до 95%, з — с 3,5 до 97,8%, з' — с 6,5 до 99%, т' — с 24,5 до 70%, ц — с 30 до 79%. Средняя разборчивость звуков в этом случае почти достигала значения, имеющего место при восприятии обычной речи с передаваемой полосой частот 2,7 кГц (значительно хуже воспринимаются после клипирования твердые взрывные). Таким образом, устранение амплитудной составляющей речевого сигнала в некоторой мере компенсируется

Таблица 7

Разборчивость начальных согласных клипированной речи в процентах, ФНЧ 7 кГц

Передано	Принято правильно	Наиболее частые замены				Передано	Принято правильно	Наиболее частые замены			
<i>б</i>	48,5	<i>м</i> —21	<i>в</i> —15	<i>л</i> —6		<i>б'</i>	38,1	<i>м</i> —14	<i>г</i> —11,7	<i>—</i> 10	
<i>д</i>	20	<i>б</i> —48	<i>в</i> —15	<i>м</i> —5		<i>д'</i>	95	<i>ж</i> —5	<i>—</i>	<i>—</i>	
<i>г</i>	42	<i>б</i> —15	<i>л</i> —11,6	<i>в</i> —9,2							
<i>п</i>	33	<i>—</i> 41	<i>к</i> —24	<i>т</i> —2		<i>п'</i>	75	<i>—</i> 13,2	<i>к</i> —6,6	<i>—</i>	
<i>т</i>	13	<i>к</i> —41	<i>п</i> —15	<i>—</i> 28		<i>т'</i>	70	<i>с</i> —13,2	<i>к</i> —10	<i>ч</i> —3,4	
<i>к</i>	75	<i>—</i> 19,8	<i>п</i> —3,6								
<i>в</i>	79	<i>—</i> 10	<i>л</i> —6,6	<i>к</i> —2,2		<i>в'</i>	54	<i>—</i> 18,4	<i>р</i> —15,2	<i>л</i> —4	
<i>ж</i>	98,3	<i>—</i> 0,9	<i>с</i> —0,8								
<i>з</i>	97,8	<i>—</i> 2,2				<i>з'</i>	99	<i>в</i> —1	<i>—</i>	<i>—</i>	
<i>ф</i>	24	<i>х</i> —30	<i>ш</i> —22	<i>ц</i> —15		<i>ф'</i>	35	<i>х</i> —20,4	<i>ш</i> —18,4	<i>ч</i> —11,2	
<i>ш</i>	97,5	<i>х</i> —0,9	<i>с</i> —0,8								
<i>с</i>	82	<i>ц</i> —15	<i>—</i> 1,6	<i>т</i> —1,5		<i>с'</i>	100	<i>—</i>	<i>—</i>	<i>—</i>	
<i>х</i>	94	<i>ч</i> —1,7	<i>с</i> —1,7	<i>ш</i> —0,9							
<i>м</i>	77	<i>—</i> 9	<i>н</i> —6	<i>л</i> —3,2		<i>м'</i>	40	<i>н</i> —32	<i>—</i> 18	<i>п</i> —5	
<i>н</i>	40	<i>м</i> —52	<i>л</i> —4	<i>—</i> 4		<i>н'</i>	43,2	<i>м</i> —40	<i>—</i> 4,2	<i>л</i> —2,8	
<i>л</i>	80	<i>р</i> —10	<i>в</i> —6,7	<i>—</i> 3,3		<i>л'</i>	61	<i>р</i> —9,4	<i>—</i> 9,3	<i>м</i> —7,4	
						<i>й</i>	90	<i>—</i>	<i>—</i>	<i>—</i>	
<i>р</i>	90	<i>л</i> —4	<i>—</i> 4	<i>в</i> —2		<i>р'</i>	94	<i>г</i> —3	<i>в</i> —1,6	<i>—</i> 1,4	
<i>ц</i>	79	<i>с</i> —18	<i>—</i> 2	<i>ч</i> —1		<i>ч</i>	91	<i>к</i> —5,2	<i>м</i> —1,8	<i>с</i> —1	

расширением передаваемой полосы частот. Дальнейшее улучшение разборчивости большинства согласных клипированной речи, в особенности твердых взрывных, *ф*, *н*, достигается ослаблением нижних частот сигнала до ограничения (табл. 8). Разборчивость конечных согласных (в слогах типа ГС) приведена в табл. 9. Эти данные следует считать приближенными, так как слоги типа ГС содержались в таблицах в меньшем количестве. В слогах ГС имелись глухие и звонкие согласные, хотя в русском языке звонкие шумные звуки в конце слов обычно произносятся как глухие. Это обстоятельство отразилось в табл. 9. Вместо звонких согласных нередко воспринимались соответствующие глухие, и наоборот. Подобные замены встречались весьма редко при прослушивании слогов типа СГ.

Анализируя данные по более частым заменам, приведенные в табл. 7, 8, 9, можно выявить некоторые особенности восприятия фонем клипированной речи. При передаче без подавления нижних

Таблица 8

Разборчивость начальных согласных клиппированной речи в процентах, ФНЧ 7 кГц, с подавлением НЧ

Передано	Принято правильно, %	Наиболее частые замены				Передано	Принято правильно, %	Наиболее частые замены			
б	63,8	г—17	м—16,6	л—2,6		б'	79,2	м—10,1	в—5,3	д—2,8	
д	89,4	г—8,4	—	—		д'	98,7	б—1,3	—	—	
г	84,4	б—6,7	л—3,1	в—2,1		г'	56,6	д—13	б—19,6	м—4,4	
п	72,2	—13,1	к—4,4	б—2,0		п'	94,7	ф—2	—1,8	в—1	
т	75,2	к—20,5	—3,2	—		т'	84,1	к—6,9	ч—4,0	—2,0	
к	81,8	—17,1	—	—		к'	89,3	т—6,5	—1,5	п—1,3	
в	76,8	л—7,4	г—4,7	—2,3		в'	83	р—5,7	л—5,2	з—3	
ж	100	—	—	—		—	—	—	—	—	
з	100	—	—	—		з'	99	—1	—	—	
ф	77,5	ш—13,6	х—7,8	—		ф'	77,5	х—8,4	ч—6,6	—3,2	
ш	92,8	ц—3,9	ж—3	—							
с	69,3	ц—29,5	—	—		с'	96,7	т—2	—1,2	—	
х	100	—	—	—		х'	58,6	ф—26	ш—3	к—2,7	
м	90,8	н—4,5	в—2,5	—		м'	73,7	н—24,9	—	—	
н	72,3	м—19,4	—3,1	—		н'	81,7	м—13,1	в—2	—1	
л	98,9	—	—	—		л'	72,2	р—12	в—7,7	—3	
—						j	99,2	—0,7			
р	82,9	л—5	г—4,7	ц—1,6		р'	100				
ц	78,0	с—22				ч	73,4	к—13,6	х—3,9	м—3,8	

частот принятые неправильно начальные твердые согласные в большинстве случаев воспринимались как губные согласные. Неправильно принятые конечные согласные записывались большей частью как фонемы с тем же местом образования (в том числе и при подавлении нижних частот).

Интересно проследить за влиянием замены после клиппирования плавного нарастания щелевых на быстрое, свойственное аффрикатам. Фонемы *ш* и *ч* друг в друга не переходили, но начальные *с* в значительном числе случаев воспринимались как *ц*, а *ц* — как *с*. Восприятие части начальных *с* как *ц* понятно. Переход многих начальных и в особенности конечных *ц* в *с* объясняется, по-видимому, приспособлением слушающих к резкому нарастанию сигнала для всех фонем, в связи с чем характер нарастания начинает терять свою различительную роль. Конечные *с* в *ц* не переходят ввиду отсутствия глухой смычки между гласным и *с*.

Разборчивость конечных согласных клипированной речи в процентах, ФНЧ 7 кгц

Фонема ϕ характеризуется значительно меньшей интенсивностью, чем c и $ш$ (в среднем на 12,5 дБ [225]). Роль признака интенсивности оказалась весьма существенной при отсутствии ослабления нижних частот до клиппирования. Большинство начальных ϕ и значительная часть конечных ϕ воспринимаются в этом случае как x и $ш$. При ослаблении нижних частот возрастает роль различий между ϕ и этими щелевыми по спектру, в результате чего 78,6% ϕ принимаются правильно.

27

В недавней работе Гупта и др. [160] исследовалась разборчивость согласных клиппированной речи для языка хинди. Речевой материал состоял из 870 слогов, содержащих 29 начальных и 31 конечных согласных в сочетании с 10 гласными. Слова произносились мужчиной и женщиной. В прослушивании участвовали 6 лиц. Результаты исследования (выявление фонем с лучшей разборчивостью, типичные ошибки при восприятии, разборчивость отдельных фонем) в значительной мере близки к изложенным выше. Можно отметить одно из различий результатов обоих исследований. В [160] ошибочно воспринятые p и k , b и d записывались большей частью как t и соответственно d . В нашем исследовании n и k переходили большей частью друг в друга, g воспринималось как b , а b — как другие звонкие губные согласные.

§ 3. О соотношении спектров речевого сигнала и его мгновенной частоты

В основу некоторых предложенных систем компрессии или анализа речевых сигналов [85, 189, 191] было положено предположение о том, что спектр мгновенной частоты речевого сигнала существенно уже спектра речи. В связи в этом представлял интерес исследовать свойства речевого сигнала, рассматривая его как сложно-модулированное по амплитуде и частоте колебание.

В работе Марку и Дагэ [189] речевой сигнал представлен в виде

$$s(t) = a(t) \cos \varphi(t),$$

где $a(t)$ — мгновенная амплитуда; $\varphi(t)$ — мгновенная фаза; $d\varphi(t)/dt = \omega(t)$ — мгновенная частота.

Обозначим через $s_1(t)$ сигнал, клиппированный по высокой частоте с использованием несущей ω_1

$$s_1(t) = \cos [\omega_1(t) + \varphi(t)],$$

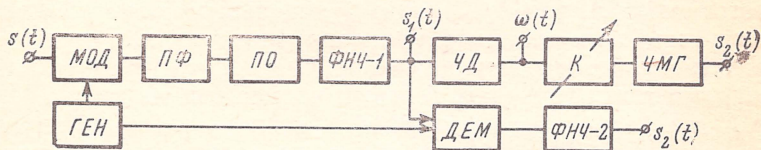
и через $s_2(t)$ тот же сигнал, перенесенный в область звуковых частот

$$s_2(t) = \cos \varphi(t).$$

Если предположить, что $\omega(t)$ меняется значительно медленней, чем $s(t)$, и на малую величину по сравнению с ω_1 , то $s_1(t)$ будет представлять собой фазомодулированный сигнал с большим индексом модуляции. Спектр этого сигнала на выходе делителя частоты в n раз должен быть в некоторых пределах уже первоначального также в n раз. О ширине спектра $s_1(t)$ после делителя частоты можно было бы судить, ограничивая спектр соответствующими фильтрами и проверяя получаемую разборчивость речи, восстановленной умножителем частоты и детектором. Однако при фильтрах с крутым срезом, обычно применяемых в многоканальной связи,

разборчивость речи резко ухудшается вплоть до полного ее пропададения, что объяснялось фазовыми искажениями в фильтрах [189]. Подобное явление отмечалось также Фишером при делении частоты $s_1(t)$ в 4 и 8 раз [145]. Вопрос о возможности использования деления частоты $s_1(t)$ или $\omega(t)$ для компрессии или анализа речи оставался неясным.

В целях исследования характера $\omega(t)$ и ее роли в переносе речевой информации был проведен эксперимент по выделению $\omega(t)$ с последующим восстановлением по ней клиппированной речи $s_2(t)$. Блок-схема установки приведена на рис. 2. Речевой сигнал $s(t)$ переносится с помощью балансной модуляции в область высоких частот с последующим выделением нижней боковой полосы.



Р и с. 2.

Блок-схема установки для выделения $\omega(t)$ с последующим восстановлением по ней клиппированной речи $s_2(t)$

МОД — модулятор; ПФ — полосовой фильтр 88,45—91,85 кГц; ПО — пиковый ограничитель; ФНЧ-1 — фильтр нижних частот 130 кГц; ЧД — частотный дискриминатор с детектором; К — каналы 1—6; ЧМГ — частотномодулированный генератор; ГЕН — генератор 92,1 кГц; ДЕМ — демодулятор; ФНЧ-2 — фильтр нижних частот 4 кГц

После ограничителя и фильтра, задерживающего вторую и более высокие гармоники, получается сигнал $s_1(t)$, следующий далее по двум направлениям: на частотный дискриминатор и демодулятор. Сигнал на выходе частотного дискриминатора (содержащего также детектор) соответствует мгновенной частоте речи $\omega(t)$. Демодулятор и фильтр нижних частот дают возможность получить $s_2(t)$.

До восстановления речевого сигнала по $\omega(t)$ спектр последней ограничивался в одном из шести каналов связи, состоящих из фильтров нижних частот с граничными частотами 150 Гц (1 канал), 250, 400, 600, 850 и 1250 Гц (6 канал). Амплитудно-частотные и фазо-частотные характеристики каналов (фильтров) приведены на рис. 3 и 4. Восстановление речи производилось генератором звуковых частот на биениях с частотной модуляцией.

Настройка всего тракта устройства осуществлялась по синусоидальному сигналу и сводилась к получению достаточно малой разности частот входного и выходного сигналов. Эта разность составляла менее 1 Гц для частот 300 и 1000 Гц и не превышала 10 Гц в диапазоне 250—1800 Гц. Мгновенная частота $\omega(t)$ в этом случае представляет собой сигнал постоянного тока с напряжением, пропорциональным частоте сигнала на входе (до ~ 2500 Гц). Правильность

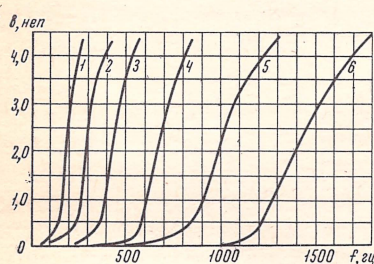
получения $\omega(t)$ проверялась более сложным сигналом

$$f(t) = a \sin 2\pi f' t + b \sin 2\pi f'' t,$$

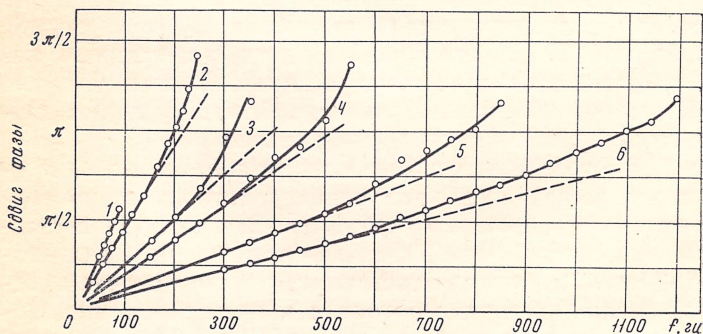
мгновенная частота которого может быть определена аналитически. При таком сигнале $\omega(t)$ представляет собой периодическую функцию с основной частотой, равной $f'' - f'$, и амплитудой, зависящей от соотношения a/b . Полученные кривые $\omega(t)$ для разных a/b соответствовали кривым, определенным аналитически.

Определение качества и разборчивости речи при воспроизведении сигнала $s_2(t)$, полученного переносом $s_1(t)$ в область звуковых частот, и $\omega(t)$ дало следующие результаты: $s_2(t)$ соответствует речи хорошего качества. При прослушивании $\omega(t)$ речь оказалась неразборчивой. Анализ кривых $s(t)$ и $\omega(t)$ для ряда слогов показал, что функции $\omega(t)$ отдельных фонем обладают довольно широким спектром, примерно таким же, как спектр исходного сигнала $s(t)$. Можно было заметить, что среднее значение $\omega(t)$ для гласных звуков большей частью близко к F_1 .

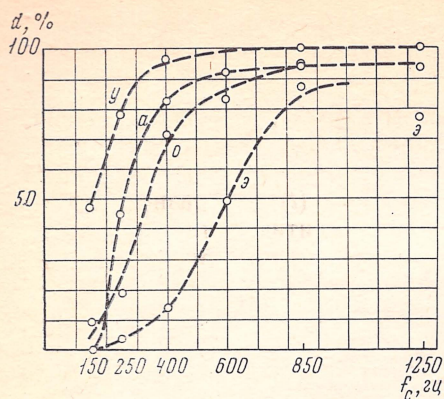
Анализ кривых $\omega(t)$, соответствующих различным звукам, не дал, однако, окончательного ответа на поставленный вопрос, так как не была исключена возможность восстановления разборчивой речи только по нижней части спектра $\omega(t)$. Поэтому было предпринято исследование речи, восстановленной по $\omega(t)$, по схеме, приведенной на рис. 2. Речевой материал состоял из изолированных



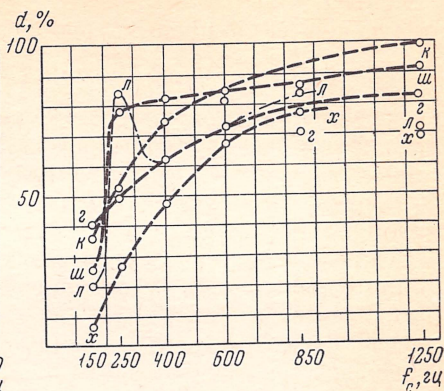
Р и с. 3



Р и с. 4.



Р и с. 5



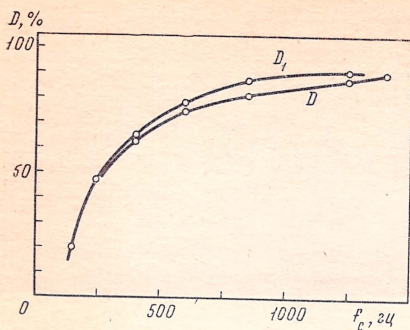
Р и с. 6

гласных, слогов типа СГ и слов. В испытаниях на разборчивость участвовали две группы в составе 12 лиц.

Звуки восстановленной речи сопровождались искажениями, характер и степень которых зависели от того, какой слог произнесен и как произнесен. При несколько ином произношении характер восстановленной речи для ряда звуков менялся в значительно большей степени. Сигнал $\omega(t)$ показал себя, таким образом, довольно неудобным для восстановления по нему речи, что подтверждает данные [189], где предпринятая попытка восстановить речь по $\omega(t)$ не удалась. Все же разборчивость ряда звуков (гласных $y, o, a, э$; согласных $к, ш, г, х, л$) при более широких каналах достигала 80—100%, что дало возможность выполнить намеченную программу исследований. Слова, состоящие из указанных выше звуков, при четком произношении также хорошо восстанавливались.

Зависимость разборчивости звуков d от ширины канала $\omega(t)$, ограничиваемого частотой f_c , приведена для гласных на рис. 5. При 5 и 6 каналах все переданные y были правильно восприняты всеми слушателями. Различия между гласными $и$ и $ы$ восстановленной речи в значительной мере терялись. Так, например, при 6 канале $и$ воспринималось правильно лишь в 37,5% случаев и как $ы$ — в 45,9%. Разборчивость $ы$ в этом канале была более высокой — 62,6%; часть остальных реализаций $ы$ воспринималась как $и$.

Кривые разборчивости d для ряда согласных приведены на рис. 6. Некоторое уменьшение разборчивости $л$ и $х$, а также $э$ при 6 канале по сравнению с 5 можно объяснить тем, что помехи, сопутствующие восстановленной речи, увеличились при расширении спектра $\omega(t)$ в большей степени, чем составляющие $\omega(t)$, важные для восстановления соответствующего звука. При прохождении $\omega(t)$ через 1 канал звуки речи приобретали «пискли-



Р и с. 7

вый» характер, а легкое постукивание вблизи микрофона воспроизводилось звуком, напоминавшим бульканье. В меньшей мере этот эффект наблюдался и при 2 канале. По-видимому, в связи с этим максимальное значение разборчивости l было получено при 2 канале. При этом канале воспринимались как l и некоторые другие звуки.

Максимальные значения разборчивости остальных согласных (за исключением c и $ц$) были в пределах 20—75 %. В значитель-

ной мере эти согласные воспринимались как близкие им звуки, отличающиеся только отсутствием или наличием голоса ($ж$ как $ш$, $б$ как $п$, $п$ как $б$) или местом образования ($т$ как $к$, $д$ как $г$, $м$ как $н$). Разборчивость $з$, $д$ и $ч$ была более низкой, а $с$ и $ц$ не были ни разу восприняты правильно.

При передаче $\omega(t)$ важно сохранить форму кривой. Для проверки влияния нелинейности фазовых характеристик фильтров на качество восстановленной речи $\omega(t)$ передавалась также через каналы, образованные двумя одинаковыми последовательно включенными фильтрами (т. е. с удвоенными фазовыми искажениями). При предварительных испытаниях ощутимой разницы в качестве восстановленной речи, при прохождении $\omega(t)$ через каналы с одинарными и удвоенными фазовыми искажениями, не было обнаружено.

Приведенные на рис. 5, 6 кривые показывают неуклонное снижение разборчивости звуков по мере сужения полосы частот $\omega(t)$. Такая же тенденция имеет место и для звуков, восстановленных с пониженной разборчивостью. Данные о разборчивости звуков восстановленной речи в целом могут быть основаны на нескольких звуках, достаточно хорошо воспринятых при прохождении $\omega(t)$ че-

рез 5 или 6 каналы. Поэтому кривая $D_1 = \frac{1}{n} \sum_{k=1}^n d_k$ при $n = 9$

(рис. 7) отражает лишь приближенно разборчивость звуков речи в целом. Для сравнения на рис. 7 приведена также кривая разборчивости звуков D для обычной речи. Для этой кривой f_c представляет собой ширину полосы пропускания фильтра. Кривые D_1 и D довольно близки друг к другу. В области ниже 500 Гц эти кривые почти совпадают. Однако выше 1000 Гц величина D продолжает расти, в то время как D_1 почти достигает предельного значения. Последнее, по-видимому, вызвано тем обстоятельством, что при более широких каналах ограничение спектра $\omega(t)$ меньше сказывается на разборчивости речи, чем помехи при восстановлении речи.

По имеющимся данным, канал шириной в 1400 *гц* (при передаче полосы 300—1700 или 400—1800 *гц*) имеет разборчивость звуков $D = 90\%$, что обеспечивает разборчивость фраз $I = 98\%$. Если принять предельное значение кривой D_1 за 100%, то $D_1 \approx 90\%$ достигается при ширине спектра $\omega(t) \cong 700$ *гц*, что следует считать несколько заниженной величиной. Следовательно, для того, чтобы восстановить по $\omega(t)$ речь с качеством, соответствующим качеству речи, переданной через канал 300—1700 *гц*, ширина спектра $\omega(t)$ должна быть более 700 *гц*.

Для получения более точных значений, а также для сравнения спектров $s(t)$ и $\omega(t)$ в условиях, соответствующих передаче речи в нормальном телефонном или широкополосном канале (где существенному улучшению качества речи сопутствует незначительное увеличение D и I), восстановленная речь не должна заметно отличаться от натуральной. Следует, однако, заметить, что от подобного рода работы, выполнение которой связано с значительными трудностями, по-видимому, не следует ожидать существенно новых данных. Результаты этого исследования были приведены в [95] и в сокращенном изложении в [50].

Исследование речи, восстановленной по $\omega(t)$, было проведено также Черри и Филлипсом [127] по схеме, близкой к приведенной на рис. 2. Разборчивость английских слов W , составлявшая 88% при ограничении спектра $\omega(t)$ фильтром НЧ с $f_c \cong 4500$ *гц* и легкой крутизной, постепенно уменьшалась до 40% при фильтре с $f_c \cong 500$ *гц* (по данным 6 слушателей). Оба указанные значения W соответствуют разборчивости звуков $D \cong 80\%$ и $D \cong 50\%$. Слушатели отмечали также неприятный характер звучания компрессирующей речи. При фильтрах с резким спадом частотной характеристики разборчивость слов была еще более низкой. Экспериментальной проверке подверглась также система передачи сигнала $s_1(t)$ с делителями частоты.

На основании приближенного рассмотрения речевого сигнала, как состоящего из двух составляющих, в работе [127] предложено использовать $\omega(t)$ для выделения формант. Для этой цели следует оставить в сигнале две соседние форманты, одна из которых должна превышать по амплитуде другую. Последующей фильтрацией $\omega(t)$ достигается удаление информации о более слабой форманте, благодаря чему восстановленный сигнал $s_2(t)$ должен быть одноформантным. Такие цепи предлагалось использовать для определения F_1 и F_3 . Значение F_2 следует при этом определить по разности $F_2 - F_1$, пропорциональной среднему числу пиков и впадин $\omega(t)$.

Экспериментальные исследования показали, что спектры речевого сигнала и его мгновенной частоты $\omega(t)$ сравнимы по ширине. Теоретическое рассмотрение ширины спектра мгновенной частоты (фазы) для ряда простых случаев (двухчастотный сигнал, белый шум) дало аналогичные результаты (Лев [58], Угер [87]).

ОПОЗНАВАНИЕ РЕЧИ НА УРОВНЯХ СЕГМЕНТОВ И ЗВУКОВ

§ 1. Признаки речевого сигнала на уровне сегментов

Многоступенчатой структуре машин, опознающих речевые сигналы на нескольких уровнях (начиная от сегментов и кончая фразами и сообщением в целом), должна соответствовать многоступенчатая система признаков. Рассмотрим низшую ступень этой системы — признаки сегментов.

О связи сегментных признаков
с дифференциальными признаками фонем

Признаки речевого сигнала на уровне сегментов должны отражать характерные свойства элементарных отрезков речевого сигнала. Признаки сегментов отождествляются часто с дифференциальными признаками фонем, понятие о которых как о фонологически существенных признаках, свойственных данному звуковому образованию (фонеме), сложилось в результате работ Трубецкого и главным образом Jakobsona, Фанта и Халле [88, 115, 161, 176, 177]. Согласно трактовке этих авторов непрерывный речевой поток подразделяется на некоторое число отрезков, имеющих различные сочетания бинарных дифференциальных признаков [176]. Каждый из этих отрезков представляет собой некоторую фонему, определяемую как связка или сочетание нескольких бинарных признаков.

Дифференциальные признаки могут быть описаны как артикуляционными терминами, так и акустическими. Трубецкой указывал, что для фонолога совершенно безразлично, какой из указанных терминологий он пользуется [115]. Для техника это далеко не безразлично, так как, по крайней мере на уровне сегментов, обе характеристики не эквивалентны и связь между ними весьма сложна (наиболее яркий пример: при различных видах артикуляции глухой смычки ее акустическая характеристика одна и та же — сигнал отсутствует). Впрочем, получение подобной эквивалентности и явилось бы, по существу, решением проблемы автоматического опознавания речи, так как артикуляционные характеристики (степень раствора; положение языка, губ, мягкого нёба; состояние голосовых связок) однозначно определяют произнесенный звук.

Применение системы дифференциальных признаков в неизменном виде для автоматического опознавания речи наталкивается на значительные трудности. Наряду с неоспоримыми достоинствами

обнаруживаются недостатки этой системы, если рассматривать ее исходя из реальной речи (того, что произносят), а не с позиции абстрактных фонем (того, что намереваются произнести).

При построении системы дифференциальных признаков за элементарную единицу речи была выбрана фонема, так как она является наименьшей единицей, выполняющей смыслоразличительную роль. В речевом сигнале, однако, фонема отображается последовательностью нескольких, порой довольно разнородных, сегментов. Крайние из них (переходные) характеризуют одновременно обе соседние фонемы, к любой из которых они принципиально могут быть отнесены. Таким образом, минимальной «звукоразличительной» единицей в речевом сигнале является сегмент. Поэтому при решении задачи автоматического опознавания речи целесообразно считать элементарными сегментные признаки и определяемые по ним сегменты. Фонемы должны опознаваться по фонемным признакам, характеризующим в общем случае некоторую определенную последовательность сегментов.

Другим недостатком системы дифференциальных признаков, вызвавшим ряд критических замечаний, является ее дихотомичность (см., например, [6, 35] с оговоркой в отношении части признаков [102]). Однако при рассмотрении этого вопроса необходимо учитывать, что отражает признак — способ или место образования звука. Для признаков, характеризующих группы звуков, одинаковых по способу образования, дихотомичность системы оправдана. Действительно, звонкость, шумность или назальность, имеющие место при действии голосового и (или) соответственно шумового источника или при отсутствии преграды между носовой и ротовой полостями, по самой природе своей характеризуются двоичным образом. Однако при их выделении из речевого сигнала с помощью различного рода цепей отсутствию того или иного признака обычно соответствует некоторое ненулевое значение; наличие признака в этих случаях устанавливается по превышению им определенного порогового значения.

Наличие или отсутствие признака звонкости у некоторой фонемы не следует связывать с обязательным наличием противопоставлений в соответствующих минимальных парах. Так, например, фонема *л* обладает звонкостью, а *ц* — не обладает, несмотря на отсутствие в русском языке глухого варианта первой и звонкого варианта второй. Логично, подходя с инженерной точки зрения, нельзя согласиться с высказыванием Бондарко и Зиндера [6] о том, что звонкость не является двоичным признаком, так как фонемы минимальных пар (например, *с* и *з*) различаются не только по наличию или отсутствию голоса, но и по другим диапазонам спектра. Наличие или отсутствие звонкости не должно означать, что во всем остальном фонемы, составляющие минимальную пару, полностью схожи. Попутно заметим, что и без голоса (в шепотном произношении) фонемы *ф* и *в* имеют существенные спектральные различия.

— Что касается признаков, определяющих место образования звуков, таких, например, как компактность или низкая тональность, то они могут выражаться различными числовыми значениями и явно не укладываются в рамки дихотомии. Для них нет такой пороговой величины, по которой можно было бы разделить все фонемы, обладающие и не обладающие данным признаком. Различительные функции выполняются этими признаками по их относительным значениям для каждой пары противопоставляемых фонем.

Авторы системы дифференциальных признаков стремились ограничиться минимальным числом признаков, дающим возможность описать все фонемы. Ввиду этого в систему не попал ряд важных признаков, таких, как длительность сегмента, крутизна фронта нарастания сигнала, плотность нулей и другие, хотя они отчасти дублируют имеющиеся. Впрочем, практически они не столько дублируют, сколько дополняют друг друга. Например, взрывной согласный в начале слова (при неслитном произнесении с предыдущим словом) не может характеризоваться глухой смычкой, так как последняя совпадает с предшествующей слову паузой. Наличие взрывного звука в этом случае можно определять по крутизне фронта нарастания сигнала, длительности, шумности. В то же время взрывные, расположенные в конце слова, хорошо характеризуются смычкой и в меньшей мере фронтом и длительностью, так как обычно они произносятся слабее и менее отчетливо. Можно поэтому согласиться с Дукельским [35], что деление признаков на дифференциальные и избыточные не оправдано. Следует иметь просто признаки, определение которых соответствует принятому в опознавании образов [92].

Групповые признаки сегментов, одинаковых по способу образования

Признаки речевого сигнала должны изменяться в значительно меньшей степени в зависимости от особенностей произношения отдельных лиц и отчасти контекста, чем сам речевой сигнал. Не следует их считать инвариантными (по крайней мере в настоящее время), так как ориентация на инвариантность ограничивает ассортимент возможных признаков, что делает число их недостаточным.

Рассмотрим некоторые сегментные признаки, свойственные группам звуков, одинаковых по способу образования.

1. Звонкость ζ позволяет выделить звонкие звуки; характеризуется наличием периодического колебания с гармониками, кратными основному тону голоса, а также большой энергией в области низких частот (НЧ-области) e_n . Этот признак может быть выделен из сигнала по наличию основного тона и с некоторым приближением по величине e_n , или по отношению этой величины, например, к энергии сигнала в ВЧ-области e_v . Признак звонкости ζ имеется в наличии в случаях озвончения глухих согласных (например, перед

звонкими взрывными) и отсутствует при оглушении звонких согласных. В случаях частичного озвончения или оглушения согласный содержит звонкий и глухой отрезки, каждый из которых может состоять из одного или большего числа сегментов. Согласно [130, 137] признак ζ может быть выделен по асимметрии огибающих положительных и отрицательных полувольт речевого сигнала.

2. Гулкость позволяет выделить тональную смычку σ' , предшествующую звонким взрывным; характеризуется наличием звонкости при почти полном отсутствии энергии в ВЧ-области e_v .

3. Отсутствие сигнала выделяет паузы в речевом сигнале, и в том числе (в середине слов) глухие смычки σ , предшествующие глухим взрывным и аффрикатам.

4. Гласность γ (неконсонантность) характеризуется большой интенсивностью сигнала и наличием в нем формантной структуры; может быть приближенно получена по величине энергии в полосе 400—800 $гц$; позволяет выделить в первую очередь ударные гласные.

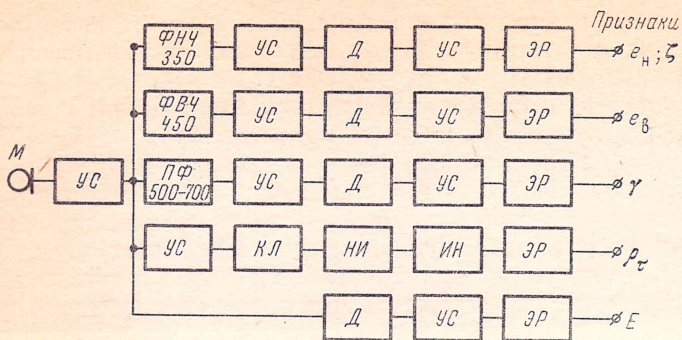
5. Шумность ρ характеризуется наличием в сигнале составляющей со сплошным спектром. Этот признак должен дать возможность выделить шумные звуки. Длительные шумные (за исключением x в сочетаниях с задними гласными) могут быть выделены по величине усредненной плотности нулей, превышающей пороговое значение. Заметим, что плотность нулей ρ_n отражает в не меньшей мере место образования звука.

6. Крутизна нарастания сигнала для шумных звуков характеризует тип смыкания в речевом тракте (является ли оно полным или неполным). Этот признак должен выделить из группы шумных глухие взрывные и аффрикаты. Тип смыкания влияет, кроме того, на длительность шумового сегмента, также являющейся групповым признаком для различения взрывных, аффрикат и щелевых.

7. Назальность ν характеризуется наличием в сигнале при носовых звуках формант, обусловленных носовой полостью (наиболее сильная на частоте около 250 $гц$).

Применение перечисленных групповых признаков должно позволить расчленить речевой сигнал на сегменты в соответствии со способом образования звуков. Однако возможности построения схем, выделяющих отдельные групповые признаки, различны. Звонкость, гулкость, плотность нулей определяют достаточно надежно соответствующие группы звуков и нашли применение во многих опознающих устройствах (см., например, [23, 102, 169, 208, 229]). Построение схем, надежно выделяющих другие признаки — крутизну нарастания сигнала, назальность, наталкивается на значительные трудности. Например, значения крутизны нарастания взрывных получаются заниженными при недостаточно четком произношении, а также при воздействии даже небольшой акустической помехи во время окончания глухой смычки.

На рис. 8 приведена блок-схема цепей, использованных в описываемых ниже исследованиях для получения групповых признаков:



Р и с. 8

Блок-схема цепей для получения групповых признаков

М — микрофон; УС — усилитель; ФНЧ — фильтр нижних частот; ФВЧ — фильтр верхних частот; ПФ — полосовой фильтр; Д — детектор и сглаживающий фильтр; КЛ — клиппер; НИ — нормализатор импульсов по длительности; ИН — интегратор; ЭР — электронное реле

звонкости, гулкости, отсутствия сигнала, гласности и усредненной плотности нулей, а также общей интенсивности сигнала. Последний признак использовался в некоторых случаях для отсеивания тихо произнесенных слов и указания на необходимость повторить их более громко. Вопросы применения групповых признаков рассмотрены в гл. 4.

Признаки сегментов, отражающие место образования звуков

Значения признаков, характеризующих место образования сегментов, непрерывно изменяются в соответствии с изменением положения активных речевых органов при речеобразовании. Изменения признаков могут быть несущественными (на протяжении квазистационарных сегментов длительных звуков — гласных, щелевых) и существенными (на протяжении переходных сегментов), где величина изменения признака является зачастую основным параметром, характеризующим сегмент. Отметим два основных типа признаков, отражающих место образования звука и, следовательно, в какой-то мере конфигурацию полостей речевого тракта.

1. Формантные частоты F_1 , F_2 являются хорошими признаками для опознавания гласных. Значения F_1 и в особенности F_2 переходных сегментов гласных содержат информацию о смягчении или месте образования соседних согласных, а также о наличии дифтонгоидной структуры самого гласного. Эти вопросы рассмотрены в §§ 3, 4 этой главы.

2. Распределение длительностей и количество интервалов между нулями клиппированной речи хорошо характеризует шум-

ные звуки (см. § 5 этой главы). В переходных сегментах переднеязычных щелевых (в сочетании с гласными) заметно увеличение длительностей интервалов между нулями, сопровождающееся уменьшением числа нулей. Степень изменения распределения указанных длительностей зависит от типа соседнего гласного и может служить признаком этого гласного.

Блок-схемы цепей для получения значений формантных частот, а также количества и длительностей интервалов между нулями описаны в §§ 3—5 при рассмотрении вопросов опознавания гласных и щелевых.

§ 2. Вопросы сегментации речевого сигнала

Под сегментацией обычно понимается расчленение речевого сигнала на наименьшие элементы — сегменты, отличающиеся друг от друга сочетанием признаков или характером изменения некоторых из них. Полное решение этой задачи позволило бы свести речевой сигнал до последовательности сегментов и опознавать по ней речь на фонежном и более высоких уровнях.

Трудности при осуществлении сегментации возникают из-за непрерывного изменения признаков, характеризующих место образования звуков как на «стыках» между сегментами (фонемами), так и на протяжении отдельных сегментов. По этой причине не удастся найти критерии для установления в общем случае границ между сегментами.

Решение задачи сегментации осуществляется по двум направлениям. С одной стороны, по части групповых признаков производится неполное членение сигнала, при котором в соответствии с используемыми признаками выявляется часть границ между сегментами [23, 82, 102, 157, 169, 208] (см. также § 3 гл. 4). Применение признака звонкости ζ позволяет расчленить сигнал на звонкие и глухие отрезки, каждый из которых состоит в общем случае из нескольких сегментов. Границы тональных смычек, звонких, щелевых и других сегментов в звонких отрезках определяются по отсутствию энергии в ВЧ-области e_v , шумности p и другим групповым признакам. Установление границ может производиться непрерывно или в дискретные моменты времени при получении отсчетов значений признаков. С помощью групповых признаков нельзя, естественно, разделить переходные и квазистационарные сегменты длительных фонем, а также соседние фонемы, принадлежащие к одной группе.

Другое направление заключается в исследовании методов сегментации в отдельных частных случаях — для некоторых определенных звукосочетаний. Так, для установления границы между двумя соседними гласными Сакаи и Дошита [208] использовали признак стабильности значений F_1 и F_2 в нескольких соседних отсчетах. Существенное изменение F_1 или (и) F_2 (при сохранении из-

менных значений на нескольких отсчетах) указывает на появление нового гласного. Границей между глухими согласными (а также звонкими взрывными) и гласными может служить первый импульс клиппированного речевого сигнала длительностью, превышающей примерно 1 мсек [99, 101]. Этот признак дублирует в известной мере звонкость. Однако он устанавливает границу точнее (обычно без запаздывания, имеющего место при использовании звонкости) и применялся в ряде наших исследований.

В работе Голда [157] звонкий отрезок, имеющий наибольшую энергию в слове, при наличии в огибающей провала (не более 0,33 одного максимума и не более 0,7 другого) расчленяется на три сегмента, представляющие соответственно гласный, звонкий согласный, гласный. Границы между ними определяются по максимуму отношения

$$\frac{\sum_{i=5}^{16} |x_i - y_i|}{\sum_{i=5}^{16} (x_i + y_i)},$$

где x_i и y_i — суммы четырех отсчетов влево и вправо от рассматриваемого момента времени в i -том спектральном канале. Применение этого правила позволило расчленить, например, звонкий отрезок *dire* (из слова *directive*) на три части: $di - r - e$.

Принцип изменения или относительной стабильности интенсивности сигнала положен также в основу метода членения сигнала на переходные и стационарные сегменты (Редди [207]). Речевой сигнал разделяется на отрезки длительностью 10 мсек. Эти отрезки считаются акустически подобными и объединяются, если интенсивность любого из них изменяется не более чем на 1/8 максимальной интенсивности соседних отрезков. Шумные отрезки объединяются по числу нулевых пересечений. За стационарный сегмент принимается объединенный отрезок длительностью не менее 30 мсек. Все остальные отрезки считаются переходными сегментами. Для примера приводим результат членения фразы *John has a book* на дискретные части и последующего их анализа — $CH\Delta A\Delta\Delta NN$ $HA\Delta ZH\Delta EH\Delta\sigma\Delta BU\Delta\sigma K$, где Δ — переходный сегмент, σ — оглушение сигнала (смычка). Эксперимент проводился с одним диктором.

Наряду с экспериментальными исследованиями по определению границ сегментов ведутся работы по построению алгоритмов оптимального членения речевого сигнала (Винцюк [12]). Задача, поставленная в этой работе, — разбить последовательность спектральных отсчетов (элементов) на сегменты, характеризующиеся минимальными изменениями спектра, направлена на выделение квазистационарных сегментов.

§ 3. Оpoznание гласных по квазистационарным сегментам

Основная информация о принадлежности гласных содержится в их квазистационарной части. Поэтому опознавание гласных обычно сводится к анализу и опознаванию их квазистационарных сегментов, без подчеркивания этого обстоятельства в большинстве случаев. Для того чтобы не нарушать эту «традицию», а главным образом для сокращения записи, в дальнейшем изложении квазистационарные сегменты гласных будут большей частью фигурировать как гласные. Заметим, однако, что при опознавании неоднородных гласных, имеющих дифтонгоидную структуру (часть реализаций *o*; *y* после губных согласных; гласные после мягких согласных, исключая *и*), следует учитывать также и начальный переходный сегмент.

Краткий обзор

Опознаванию гласных посвящено большое число работ, в подавляющей части которых для получения признаков этих звуков используется спектральный анализ речевого сигнала в диапазоне от 100—200 *гц* до 5—10 *кгц*, выполняемый обычно при помощи нескольких параллельно включенных полосовых фильтров.

Впервые распределение энергии сигнала по спектру в целях получения признаков для автоматического опознавания гласных (а также других звуков речи) исследовалось Мясниковым [66]. В его работе, как и в ряде следовавших за ней исследований [45, 135, 140, 215], в качестве признаков гласных использовались знаки разностей или произведения значений энергии в нескольких (от 5 до 32) взятых попарно полосах частот. В [140], кроме того, выделялись дополнительные спектральные признаки из производных по времени от временных огибающих в отдельных полосах частот. Надежность опознавания гласных в этих исследованиях составляла 75—80%. Речевой материал состоял из изолированных звуков, открытых слогов и небольшого числа односложных слов, произносимых двумя-тремя лицами (в [45] — 20 лицами). Надежность опознавания гласных при помощи рассмотренного метода может быть повышена, если частотные полосы будут ближе соответствовать формантным областям гласных, как это показал Варшавский [8]. По максимуму энергии в области первой форманты (диапазоны 200—400, 400—700, 700—1100 *гц*) звуки *y*, *o*, *a* различались с надежностью ~ 90%. Гласные выделялись из ключевых слов (по одному слову на каждый звук), произносимых 60 дикторами.

Другим направлением исследований по опознаванию гласных явилось использование в качестве признаков гласных значений формантных частот. В работе Петерсона и Барни [202] в качестве наиболее информативных признаков гласных принимались первые три формантные частоты: F_1 , F_2 и F_3 , их амплитуды и частота основного тона F_0 . Исследовались 10 английских гласных, произне-

сенных в односложных словах типа h — гласный — d по два раза 76 дикторами обоего пола (взрослыми и детьми). Криволинейные области (большей частью выпуклые), проведенные на плоскости F_1F_2 , охватывали до 90% соответствующих гласных. В результате прослушивания этих слов 26 компетентными лицами были признаны типично звучащими 1199 гласных из общего числа 1520. Данные [202] по типично звучащим гласным использовались в качестве исходного речевого материала в ряде последующих работ. Так, Фаулкис [154], преобразовав значения F_1, F_2, F_3, F_0 в три параметра, получил прямоугольные области вместо криволинейных при средней надежности опознавания гласных 88,5%. В [228] для этих же исходных данных (за исключением гласных, произнесенных детьми) применены два многомерных статистических метода решения, примерно равноценных по результатам действия. При помощи ЦВМ IBM704 была достигнута надежность опознавания 90,7% при использовании в качестве признаков F_1, F_2 и F_3 и 94,4% — при добавлении к ним F_0 и отношений амплитуд формантных частот. После попытки получения собственных данных о формантах у авторов [228] сложилось мнение, что автоматизация процесса определения формантных частот с высокой точностью исключительно трудна.

Исследования по опознаванию английских гласных по формантным частотам и соотношениям энергии в определенных частотных полосах при помощи ЦВМ проводились Форджи [152]. Исследовались 10 гласных, 9 из которых произносились в словах типа b — гласный — t . Классификация отрезков сигнала длительностью 5,5 мсек начиналась с грубого определения F_1 и F_2 , в результате чего каждый гласный попадал в одну из 11 прямоугольных областей на плоскости F_1F_2 , содержащих от одного до шести звуков. Разделение последних осуществлялось по соотношениям энергий определенных частотных полос с учетом информации об основном тоне. Средняя надежность опознавания гласных при 24 дикторах обоего пола, произносивших испытательные слова по одному разу, составила 88%, а с учетом длительности звуков — 93%.

По формантным частотам, по их отношению и сумме могут определяться так называемые дифференциальные признаки, по которым может быть принято решение при опознавании гласных. В работе Судзуки и Наката [221] японские гласные разделяются по признаку F_2/F_1 на диффузные — i , компактные — a, o и недиффузные, некомпактные — e, u . Разделение последних двух групп на простые и бемольные, высокотональные и низкотональные осуществлялось по F_1 и F_2 . Речевой сигнал анализировался в диапазоне 200—5900 $гц$ при помощи 26 полосовых фильтров и после преобразования в двоичный код поступал на ЦВМ. Формантные частоты определялись методом вычисления моментов с точностью до $\pm 5\%$ для F_1 и $\pm 10\%$ для F_2 при $F_0 < 200$ $гц$. Средняя надежность опознавания гласных составила 98% при участии в эксперименте двух

дикторов — мужчин, произносивших по 300 слогов. Оpozнaвание десяти английских гласных по вычисленным на ВМ IBM 7090 значениям дифференциальных признаков проводили Хемдал и Хьюз [169]. В качестве речевого материала применялись 227 закрытых слогов, 50 односложных слов и отрезки непрерывной речи. В отличие от [221] разделение гласных на компактные и диффузные производилось по F_1 , а на простые и бемольные — по $F_1 + F_2$. Кроме того, по длительности и сдвигу F_1 и F_2 различались напряженные и ненапряженные гласные. Надежность опознания гласных в словах для одного диктора составила 92% и приблизительно соответствовала результатам восприятия этих слов группой слушателей. Для поддержания полученной надежности опознания для двух других лиц оказалось необходимым изменить граничные пороговые значения признаков.

Основанный на статистической теории обнаружения сигналов метод опознания гласных, при котором огибающая спектра звука сравнивается со средними спектрами гласных, можно считать третьим направлением исследований в этой области. В работе Чистович [105] описаны метод опознания и средние спектры русских гласных, произнесенных в шести ключевых словах 40 лицами. Каждая спектрограмма представлялась в виде последовательности из 11 значений огибающей спектра. Методом определения вероятностей гипотез была получена надежность опознания гласных, составлявшая 86%. По речевым данным, полученным после анализа сигнала 35 полосовыми фильтрами Форджи [152], Смит и Клем [216] получили средние спектры 10 английских гласных. Путем применения квадратичной классификационной функции к ковариантным матрицам, вычисленным для каждого гласного, им удалось повысить полученную Форджи надежность опознания с 88 до 94% (без учета длительности звуков в обоих случаях). Судзуки и Ондзуми [223, 224] выразили огибающую спектра каждой из 5 японских гласных десятизначным двоичным кодом (после анализа 10 фильтрами в диапазоне 250 гц — 7 кгц). Классификационная функция определялась из условия максимума отношения квадрата суммы разностей признаков двух различаемых гласных к дисперсии этих гласных. Надежность опознания изолированных гласных составила 96% при 138 дикторах обоего пола и 665 реализациях. Оpozнaвание 5 итальянских гласных вероятностным методом (после анализа сигнала 16 фильтрами в диапазоне 100 гц — 5 кгц) описано Фавелли и др. [150]. Обучение производилось по 10 реализациям двух мужчин. Надежность опознания при 50 реализациях составила 92%.

К четвертому направлению исследований в рассматриваемой области можно отнести работы, основанные на анализе клиппированного речевого сигнала. Оpozнaвание гласных по клиппированному речевому сигналу основано в значительной мере на определении параметров, близких к формантным частотам. Допущение

Чанга и др. [126] о близком соответствии усредненного числа нулей речевого сигнала значению F_1 , а числа нулей после дифференцирования — F_2 в эксперименте Петерсона [201] не подтвердилось. Он же показал на небольшом речевом материале, что при разделении сигнала до клиппирования на две полосы 0,2—1 кГц и 1—4 кГц усредненное число нулей будет в первой полосе близко совпадать с F_1 , а во второй полосе весьма приближенно соответствовать F_2 .

Анализ пяти японских гласных по клиппированному речевому сигналу описан Сакаи и Иноуэ [210]. Интервалы между нулями квантовались по длительности на 14 градаций. Число этих интервалов регистрировалось счетчиками отдельно для всех градаций длительности. Каждый гласный характеризовался положением одного или двух максимумов распределения временных интервалов. В работе [208] из сигнала до клиппирования выделялись области первых двух формант: 100—1500 и 700—2500 гц. Длительности интервалов между нулями квантовались на 5 градаций в области F_1 и на 3 градации в области F_2 . Надежность опознавания гласных по положению максимумов распределения длительностей интервалов в обеих областях превышала 90% для слогов типа СГ (согласный — гласный), произнесенных мужчинами. Для женских голосов должна быть изменена характеристика фильтра, выделяющего область F_2 . Надежность опознавания была затем улучшена при некотором усложнении метода принятия решения до 94% для мужских голосов (по 11 дикторам, 1145 реализациям гласных) и 90% для женщин (9 дикторов, 840 реализаций) (Дошита [138]).

Бездэл и Чандлер [120] исследовали 5 английских гласных (а, е, ои, i, u), произносимых в ключевых словах. Анализ речевого сигнала производился в диапазоне 300—3400 гц по 16 градациям длительностей интервалов между нулями, а опознавание — по 6 укрупненным градациям. Надежность опознавания для 10 дикторов, по которым устанавливались признаки, составила 88% — при принятии решения по минимуму евклидовых расстояний и 94% — при добавлении весовых коэффициентов, равных обратным значениям среднеквадратичного отклонения. Для 40 других дикторов надежность опознавания гласных составила 79% при решении первым из этих методов. В последующей работе Бездэл анализирует речевой сигнал отдельно в двух диапазонах частот — ниже и выше 1000 гц, разделяя длительности интервалов на 5 и 8 градаций.

Как следует из краткого обзора, в большинстве опубликованных работ (более подробные данные о некоторых из них приведены в [59, 78, 100]) речевой материал состоит из гласных, произносимых изолированно или в односложных ключевых словах, что лишь приближенно отражает разнообразие характера ударных гласных в различных звуко сочетаниях и многосложных словах. При сопоставлении результатов разных исследований необходимо также

учитывать наличие (или отсутствие) других факторов, уменьшающих разброс исходных речевых данных (например, участие малого числа дикторов, использование только дикторов-мужчин, отбрасывание нетипичных реализаций при предварительном прослушивании). Наиболее информативными признаками являются F_1 и F_2 . Однако вычисление этих признаков связано с рядом трудностей [8, 78, 152, 228] и во многих работах выполнялось не в реальном масштабе времени.

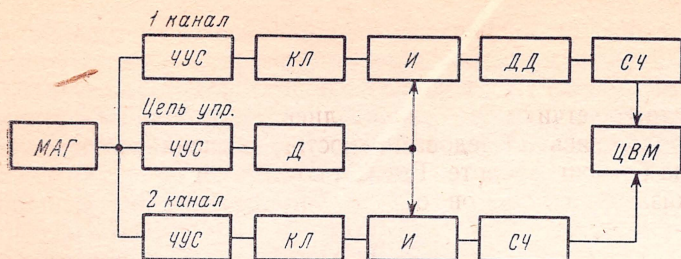
Целью настоящего исследования являлось получение небольшого числа весьма информативных признаков гласных сравнительно простым способом для применения их при опознавании наборов слов. Поставленная задача привела к необходимости выделения признаков, близких к формантным данным, по клиппированному речевому сигналу, при использовании сравнительно разнообразного речевого материала.

Исследование признаков квазистационарных сегментов гласных

Анализ гласных и получение признаков для их опознавания производились по клиппированному речевому сигналу. При передаче в полосе частот 100—2700 гц клиппирование ухудшает разборчивость гласных y , $и$, $ы$, в особенности для женских голосов. При расширении полосы частот до 7 кгц подобного рода ухудшение почти не имеет места, что следует из табл. 4 (гл. 2), где были приведены значения разборчивости гласных для этого случая при некотором ослаблении нижних частот.

Учитывая физическую природу гласных, характеризующих в первую очередь формантными частотами, следует полагать, что разделение сигнала до клиппирования на частотные полосы, соответствующие отдельным формантам, позволит получить более информативные признаки. Однако ввиду большого разброса формантных частот для разных гласных и разных дикторов разделение на формантные области может быть выполнено лишь приближенно. Дальнейшее увеличение числа частотных полос до клиппирования может лишь ухудшить информативность признаков из-за уменьшения значимости формантных частот и возрастания роли слабых составляющих сигнала.

Ввиду трудности разделения областей F_2 и F_3 , а также малой информативности F_3 по сравнению с F_1 и F_2 было решено выделить признаки гласных по двум каналам, в одном из которых до клиппирования в большей степени усиливается область F_1 и другом — область F_2 для звуков $а$, $и$, $ы$. В отличие от [201, 208], через каждый из обоих каналов проходил весь спектр сигнала с ослаблением частот, удаленных от соответствующих формантных областей, на 6 дб на октаву. Предварительными исследованиями было установ-



Р и с. 9

Блок-схема устройства, выделяющего признаки гласных

МАГ — магнитофон типа МЭЗ-28А; ЧУС — частотнозависимый усилитель; КЛ — клиппер; ДД — дискриминатор импульсов по длительности; СЧ — счетчик; Д — детектор и сглаживающий фильтр; ЦВМ — вычислительная машина типа БЭСМ-3М

лено, что в первом приближении достаточно иметь по одной градации длительностей интервалов между нулями в каждом канале.

Блок-схема устройства, выделяющего признаки гласных, приведена на рис. 9. Речевой сигнал с магнитофона разветвляется по трем направлениям: в цепь управления и два канала выделения признаков. Усилители первого канала и цепи управления имеют спад усиления верхних частот начиная с 700 *гц*. Усилитель второго канала имеет спад усиления нижних частот начиная с 1000 *гц* и верхних — с 3000 *гц*. Дискриминатор первого канала задерживает импульсы длительностью менее 0,22 *мсек*. Элементы И пропускают сигналы при наличии достаточного напряжения в цепи управления, что имеет место при произнесении звонких звуков. Наличие сглаживающего фильтра в детекторе цепи управления приводило к запаздыванию включения элементов И примерно на 20 *мсек*. Показания счетчиков передаются непосредственно в вычислительную машину БЭСМ-3. Схема ввода признаков гласных в машину описана в [22].

В качестве признаков гласных были приняты $F_1 \cong \Delta F_1 / \Delta t$ и $F_2 \cong \Delta F_2 / \Delta t$, где ΔF_1 и ΔF_2 — число интервалов между нулями (соответствующих положительным импульсам сигнала) на выходах первого и второго каналов за время Δt .

Для проведения исследования гласных были составлены два перечня слов. В первый из них (основной) вошли 40 слов, в том числе 10 цифр и некоторые математические термины. Второй (дополнительный) перечень содержал 50 слов, состоящих из часто употребляемых слогов и сочетаний гласных с согласными, которые не вошли в первый перечень. Сведения о гласных и числе слогов в использованных словах приведены в табл. 10. Из числа гласных, расположенных после палатализованных согласных (именуемых в дальнейшем для краткости мягкими гласными), кроме *и*, исследовались чаще употребляемые *'э* и *'а*, а также, в одном слове, *'у*.

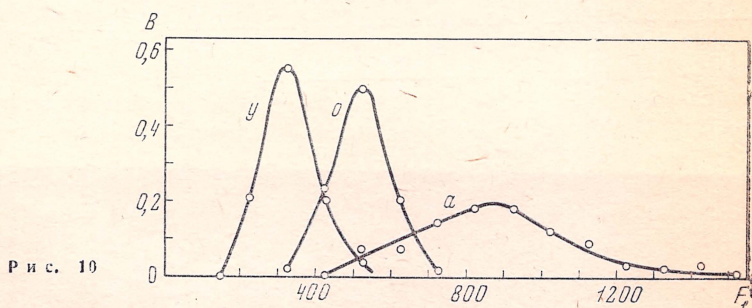
Слова первого перечня произносили 25 лиц (13 мужчин, 12 женщин), слова второго перечня — пять лиц. Показания счетчиков считывались каждые 20 мсек и выводились на печать ЦВМ. После каждого считывания счетчики устанавливались в нулевое положение. Гласные отображались последовательностью значений ΔF_1 и ΔF_2 , печатаемых на бумажной ленте. Глухим согласным соответствовали нулевые показания счетчиков обоих каналов, поэтому границы между этими согласными и гласными были на ленте достаточно четкими. Границы между гласными и звонкими согласными в подавляющем большинстве случаев были хорошо заметны. В небольшом числе реализаций положение границы было неясным, и эти реализации исключены из дальнейшего рассмотрения.

Т а б л и ц а 10

Количество гласных и число слогов в использованных словах

	Количество ударных гласных									Количество слов с числом слогов			
	у	о	а	э	и	ы	'э	'а	'у	1	2	3	4
I список . .	3	8	5	2	8	4	6	3	1	9	18	11	2
II список	8	11	15	3	6	3	—	3	—	4	33	12	1

Как и следовало ожидать, величины ΔF_1 и в особенности ΔF_2 изменялись в течение длительности звука. При предварительном рассмотрении наиболее типичным оказался третий отрезок гласного, начинавшийся примерно через 60 мсек после начала звука. По данным этого отрезка для реализаций твердых гласных (гласных, расположенных после непалатализованных согласных) и звука *и* были построены кривые плотности распределения (частоты) признаков $B(F_1)$ и $B(F_2)$, приведенные на рис. 10—13. На оси



ординат отложено отношение числа реализаций, имеющих значения $F_1 \pm 50$ или $F_2 \pm 100$, к общему числу реализаций данного звука. Кривые $B(F_1)$ разделяют гласные на три области: *уиы*, *оэ* и *а*. Положения максимумов $B(F_1)$ (равные для *у* — 320, *о* и *э* — 520, *и* — 270, *ы* — 300, *а* — 870) хорошо совпадают с величинами F_1 для этих гласных. Граничные значения $F_1 = 420$ и $F_1 = 670$, разделяющие области гласных на рис. 10, хорошо совпадают с соответствующими граничными частотами 410 и 680 *гц*, полученными из кривых статистического распределения частот формант, приведенных Варшавским [8]. Аналогичные границы получены также при исследовании восприятия синтетических гласных. Так, например, значение F_1 , разделяющее области восприятия *уиы* и *оэ*, по данным Мушников и Чистович составляет в среднем 414 *гц*. К этой величине близки также данные ряда зарубежных исследований [110].

Кривые $B(F_2)$ разделяют гласные на две области: *уо* и *аиыэ*. Положения максимумов этих кривых лишь весьма приблизительно соответствуют значениям F_2 . Это объясняется более сложной структурой сигнала во втором канале и большей зависимостью ΔF_2 (по сравнению с ΔF_1) от соотношений амплитуд формант и уровня сигнала.

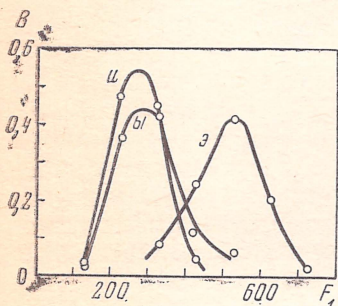


Рис. 11

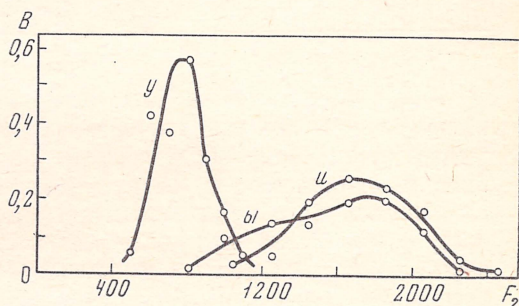


Рис. 12

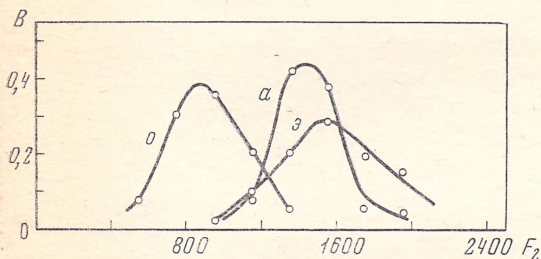
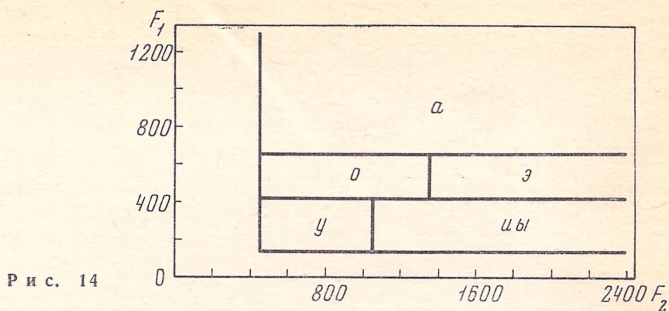


Рис. 13



Р и с. 14

Кривые $B(F_1)$ и $B(F_2)$ дают возможность разделить множество гласных на отдельные звуки, за исключением *и* и *ы*. Области распределения гласных на плоскости F_1F_2 приведены на рис. 14. По признакам одного отрезка квазистационарного сегмента эти области правильно классифицируют 70,4% звука *у*, 93% *о*, 83,2% *а*, 73,9% *э*, 92,2% *и* и 80,5% *ы*, причем последние два звука представлены одной областью и между собой не различаются. Границы между областями не всегда удастся провести оптимальным образом, так как F_1 и F_2 кратны $1/\Delta t$, что дало, в частности, пониженные данные для *у*.

Опознавание гласных по последовательности отрезков квазистационарных сегментов

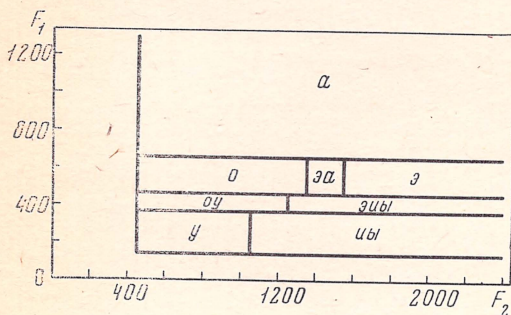
Надежность опознавания гласных может быть улучшена дополнительным выделением на плоскости F_1F_2 промежуточных областей *оу*, *эа* и *эиы* (рис. 15) и использованием признаков последовательности отрезков гласного. Для этой цели был разработан следующий алгоритм.

1. Устанавливалась принадлежность каждого отрезка в соответствии с областями, приведенными на рис. 15.
2. Упразднялись промежуточные области. Отрезки *оу* классифицировались как *у*, если рядом располагался отрезок *у*, или как *о* — во всех других случаях. Аналогично с преимуществом в пользу *э* классифицировались отрезки *эа* и *эиы*.
3. Зачеркивались одиночные отрезки, за исключением начального *и*-образного отрезка.
4. В случае, если гласный состоял из отрезков одного типа, он классифицировался в соответствии с этими отрезками. В остальных случаях гласный обозначался сочетанием составляющих его отрезков. Как правило, подобные неоднородные гласные классифицировались по более длительной части (по большинству отрезков), а в случае равной длительности — по начальной части. Из этого правила были отдельные исключения.

Надежность опознавания гласных по данным последовательности отрезков приведена в табл. 11. По сравнению с надежностью, полученной по данным одного отрезка гласного, опознавание гласного *у* улучшилось на 13,6%, *о* ухудшилось на 1,4%, опознавание остальных гласных улучшилось в среднем на 3%.

Надежность опознавания гласных зависит от характера произношения отдельных лиц и типа звукосочетаний и слов. На 36% дикторов приходится 60% ошибок. Наличие повышенного числа одинаковых ошибок в некоторых определенных словах (*сумма*, *ноль*, *целых*) указывает, что причиной этого являются особенности произношения некоторых слов и сдвиг формантных частот гласного, зависящий от соседних с ним звуков и слова в целом. Этот сдвиг обычно невелик, но достаточен для того, чтобы гласный перешел границы между областями и попал в соседнюю область. Например, в 23 реализациях слова *сумма* звук *у* в 7 случаях опознавался как *о* (в то время как в 25 реализациях слова *пуск* в *о* перешло лишь одно *у*). При добавлении правила, что после *с* (которое может быть опознано в непрерывном сигнале) промежуточная область *уо* переходит в *у*, число подобных ошибок уменьшается с 7 до 2. Большое число ошибок при опознавании *э* объясняется отчасти наличием слова *целых*, в котором на 23 реализации *э* было 9 ошибок, в том числе 6 попаданий в область *о*. Для *э*, произнесенных в четырех других словах, надежность опознавания составила 86,5%.

После отбрасывания одиночных отрезков других наименований большинство твердых вариантов гласных получилось однородным. Мягкие варианты гласных (за исключением *и*) остались в большинстве своем неоднородными. По степени неоднородности, зависящей от артикуляции, их можно разделить на 2 группы: '*э*', '*у*' и '*а*', '*о*'. Как видно из рис. 14, у гласных первой группы изменения формант при постепенном переходе от *и*-образного начального сегмента до квазистационарной части могут быть не очень большими, так как область *и* граничит с областями *э* и *у*. Гласные '*э*' и '*у*' имеют хорошо выраженный *и*-образный участок (в 85,7% случаев длительностью более 15—20 мсек, не считая длительности 20 мсек начального участка, не учитываемой вследствие запаздывания включения F_1 и F_2).



Р и с. 15

Таблица 11

Надежность опознавания гласных в процентах

Опознано	Произнесено								
	у	о	а	э	ы	и	'у	'а	'э
у	84	4,2	—	—	4	4,4	25	—	1,4
о	11	91,6	4,7	13,3	3	—	4,2	6,7	4,2
а	—	3,2	85,3	1,7	—	—	—	42,6	0,7
э	—	0,5	9,4	76,7	8	0,6	—	48	48,6
и, ы	5	0,5	0,6	8,3	85	95	70,8	2,7	45,1

Этот *и*-образный сегмент далее переходит в *э* или *у* или остается *и*-образным на большей части длительности звука. Гласному 'э в 40,3% случаев соответствует сочетание *иэ*, которое очень редко встречалось у других гласных (исключая 'а). Гласный 'у в 37,5% случаев представлен сочетанием *иу*. Однако подобное сочетание нередко встречается у *и*, в особенности до носовых согласных. Поэтому было решено мягкие гласные (сочетания *иэ*, *иу*) в графе «Опознано» табл. 11 не выделять и отнести *иэ* к *э*, *иу* к *и*. При этом гласный 'э опознается как *э* или *и* в 93,7% случаев, а 'у — как *и* или *у* в 95,8% случаев.

Изменения формант при образовании 'а более значительны, так как *и*-образный участок этого звука может переходить в квазистационарную часть только через область *э*. Гласные 'а должны были выражаться поэтому сочетанием *иэа*, но таких реализаций встретилось мало. Лишь у 30% 'а был зафиксирован *и*-образный начальный участок длительностью более 15—20 мсек. Для большинства остальных 'а этот участок получился *э*-образным. Гласный 'а опознается на 90,6% как *э* или *а*, причем в положении между двумя мягкими согласными 'а опознается преимущественно как *э*, а между мягким и твердым — как *а*. На значительную неоднородность 'а и наличие в нем *э*-образного участка указывает Дукельский [35]. В приведенных в книге Фанта [88] спектрограммах открытых слогов с гласными *а* и 'а хорошо видны большие различия между ними при согласных, различающихся только отсутствием или наличием палатализации. Длительность формантных переходов от *и* к *э* и *а* у гласного 'а в ряде случаев больше длительности *а*-образной квазистационарной части, у которой часто F_2 выше, а F_1 ниже соответствующих формантных частот твердых вариантов *а*. Это тоже указывает на несколько *э*-образный характер сигнала 'а в таких случаях. Наличие больших изменений квазистационарных сегментов 'а и 'о по сравнению с *а* и *о* отмечается также Дукельским [35].

Исследования показали, что по двум признакам гласных, получаемым сравнительно простым способом, можно опознавать твердые варианты ударных гласных и *и* со средней надежностью 88,2% (с учетом частоты встречаемости), причем *и* и *ы* входят в одну группу и не различаются между собой. Мягкие варианты ударных гласных 'э, 'а, 'у, являющиеся, по существу, дифтонгоидами, опознаются в более чем 90% случаев лишь как один из составляющих их звуков. В известной мере это уменьшает неопределенность при опознавании набора слов. Работа по лучшему представлению признаков мягких гласных (а также по различению *и* и *ы*) с использованием параметров переходных сегментов гласных изложена в §§ 4 и 7 этой главы. Добавлением двоичного признака высоты основного тона можно избежать части ошибок, вызванных чрезмерным сдвигом формантных частот при высоких женских голосах.

§ 4. Исследование переходных сегментов гласных. Признаки места образования и смягчения соседних согласных

Переходные сегменты (переходы) речевого сигнала, образующиеся «на стыках» между звуками в связи с плавными изменениями положения артикуляционных органов, характеризуют, как известно, одновременно оба соседних звука. В особенности велико значение переходов гласных, являющихся в ряде случаев важнейшими и даже единственными признаками смягчения или места образования соседних согласных. Информация о переходных сегментах гласных необходима также при опознавании неоднородных (дифтонгоидных) гласных, расположенных после мягких согласных. При расширении опознаваемого словаря от нескольких десятков до нескольких сотен слов использование признаков, характеризующих переходы гласных, становится неизбежным.

При необходимости исследования переходных сегментов гласных на большом речевом материале возникает задача быстрого, автоматического определения параметров переходов. Эта же задача, естественно, возникнет при попытке использовать эти параметры в качестве признаков согласных звуков и неоднородных гласных при опознавании набора слов. Целью настоящего исследования являлось решение части поставленной выше задачи, а именно: определить параметры переходных сегментов для сочетаний глухих согласных с гласными и звонких взрывных с гласными и получить по ним признаки места образования и смягчения соседних согласных.

О понятии «локус»

Термин «локус» (*locus*) или «положение» был введен сотрудниками Хаскинских лабораторий при исследовании переходов синтезированных гласных [133]. Под локусом понималась опорная точка

на плоскости частота — время, к которой были направлены рисованные формантные переходы гласных.

При рассмотрении натуральной речи понятию локус соответствует характерное для каждого согласного (зависящее от места его образования) значение частоты, к которому направлен переходный сегмент одной из формант соседнего гласного. Обозначим локусы первой и второй формант через $F_{1л}$ и $F_{2л}$. Локусы характеризуют предельное положение речевых органов, от которого (или к которому) активные органы начинают свое движение до положения, соответствующего квазистационарной части соседнего гласного. Начальный участок движения соответствует произнесению взрывного согласного или части щелевого. Поэтому начальные переходы F_2 лишь направлены к значению $F_{2л}$, но не достигают его. Это обстоятельство находится в соответствии с выводом сотрудников Хаскинских лабораторий, что более естественное звучание синтетических слогов имеет место, когда переход F_2 не доходит вплотную к локусу $F_{2л}$. Начальные положения формант $F_{2н}$ (или конечные $F_{2к}$) располагаются ниже локусов при нарастающих переходах (считая относительно гласного) и выше — при спадающих. При опознавании речи локусы не могут быть использованы. Они могут быть применены для проверки правильности построения динамики формантных кривых.

Краткий обзор

Рассмотрим кратко имеющиеся данные о характере и значениях параметров переходных сегментов. Для гласных русской речи характер переходных сегментов в первую очередь зависит от твердости или мягкости соседних согласных. В работе Дукельского [35] на основании прослушивания выделенных частей слогов и анализа осциллограмм отмечается, что после твердых согласных качество и длительность начального перехода определяется гласным и местом артикуляции согласного и мало зависит от способа образования согласного (за исключением назальности). Губные согласные понижают тональность перехода a (и в меньшей степени других гласных), а язычные повышают его. Начальный переход o носит большей частью y -образный характер, что указывает на несколько дифтонгоидную структуру этого гласного. Конечные переходы обладают меньшей интенсивностью и носят менее четкий характер.

Исследованиям переходных сегментов гласных, находящихся в сочетаниях с мягкими согласными, посвящены многие работы. По данным Бондарко [3], на основе анализа спектра и прослушивания слов, произнесенных 7 лицами, мягкость предшествующего согласного характеризуется наличием в гласном u -образного (в большей части $'a$ ε -образного) начального сегмента длительностью от 40 до 100 мсек. Здесь и далее добавление к обозначению гласного

штриха слева сверху указывает на мягкость предшествующего согласного. В гласных 'о, 'а начальный переходный участок может состоять из двух сегментов (*иу*, *иэ*) с суммарной длительностью порядка 100 мсек и более. Начальный переход имеет более ярко выраженную формантную структуру и большую интенсивность по сравнению с конечным. По данным Дукельского, Кузнецовой [35, 56], гласный 'а всегда начинается с *и*-образного сегмента, составляющего от 6 до 20% общей длительности звука (в зависимости от темпа произношения и диктора). В исследовании, описанном в предыдущем параграфе, хорошо выраженный *и*-образный начальный сегмент (длительностью более 30—40 мсек) был зафиксирован у 86% реализаций 'э и 'у и лишь у 30% гласных 'а.

Многочисленные исследования переходов английских гласных, основанные на данных восприятия синтезированных слогов, проводились Делаттром, Либерманом, Купером и другими сотрудниками Хаскинских лабораторий [131, 133, 134, 170, 185]. В большей части экспериментов с синтетическими слогами часть перехода гласного, прилегающая к локусу $F_{2л}$, отсутствовала. В результате этих исследований было установлено, что локус $F_{2л}$ определяется почти исключительно местом образования согласного. Начальный и конечный локусы второй форманты $F_{2л}$ для гласных в сочетаниях с губными *b*, *p*, *m*, *v*, *f* получились равными 700 *гц*, в сочетаниях с *d*, *t*, *n* — 1800 *гц*, а в сочетаниях с *z*, *s* и *ʒ*, *ʃ* — соответственно 1600 и 2000 *гц*. Лишь при взрывных *g* и *k* значения локусов зависят также от гласных. Для сочетаний этих взрывных с передними гласными *i*, *e*, *a* локус $F_{2л} = 3000$ *гц*. Для сочетаний *g*, *k* с задними гласными *u*, *o* локус значительно ниже по частоте и не поддается точному определению.

Исследования с использованием спектрограмм типа «видимая речь» показали, что закономерности переходов гласных естественной речи более сложны. Халле, Хьюз и Редли [162] анализировали произнесенные четырьмя лицами односложные слова, содержащие сочетания каждого из 10 английских гласных с взрывными в начальном и конечном положениях. Трудности возникали иногда уже при выделении переходного сегмента на спектрограмме. Одним из отличительных признаков *k* в сочетании с передними гласными является сходимость переходов F_2 и F_3 . Наименее четкими оказались переходы от глухих взрывных в начальном положении, где лишь для четырех гласных *I*, *Λ*, *U*, *u* удалось установить качественные закономерности. Начальное значение второй форманты $F_{2н}$ передних гласных в сочетаниях с *p* ниже, а в сочетаниях с *k* выше квазистационарной части (ядра) гласного. $F_{2н}$ примерно равно ядру гласного в *tI* и выше него в *tΛ*. Переходы задних гласных имеют более высокое $F_{2н}$ после *t* и более низкое — после *p* и *k*, причем оба последних переходных сегмента не различаются друг от друга. Переходы от звонких взрывных более четкие. Для всех гласных (исключая *i*, переходы которого могут быть похожими)

$F_{2н}$ и $F_{3н}$ после b ниже, чем после d . Для передних гласных наиболее высокое $F_{2н}$ наблюдается в сочетаниях с g , а для задних — в сочетаниях с d . Переходы от гласных к взрывным более однообразны. Для передних гласных конечное значение второй форманты $F_{2к}$ ниже ядра до p , b и выше или примерно равно ядру до остальных взрывных. $F_{2к}$ для g заметно выше, чем для d . Для задних гласных $F_{2к}$ выше ядра лишь при t и d . Переходы к p и k (а также к b и g) не различаются существенно друг от друга. Заметим, что в [162, 184] понятия локуса $F_{2л}$ и начального значения форманты $F_{2н}$ не различаются.

Количественные данные о переходах для естественной речи определялись в работах Лехисте и Петерсона [184], Стивенса, Хауза и Поля [220]. В первой из этих работ анализировались сочетания 15 гласных и 23 согласных в 1263 закрытых слогах, произнесенных одним лицом, и в 70 словах пяти лиц. Отмечены значительные колебания $F_{2н}$ и $F_{2к}$ для всех губных звуков, причем их среднее значение значительно превышает 700 $гц$. Например, для звука b значение $F_{2н}$ колеблется от 900 $гц$ в сочетании $b\bar{o}$ до 1780 $гц$ в сочетании bi . Для начальных переднеязычных звуков d , t , n диапазон колебаний меньше. В сочетаниях с d значения $F_{2н}$ колеблются от 1535 $гц$ в da до 1785 $гц$ в di , а $F_{2к}$ — от 1250 $гц$ (aUd) до 1960 $гц$ (id). Диапазон колебаний $F_{2н}$ и $F_{2к}$, k и g несколько превышает диапазон изменений для губных звуков и простирается от 680 $гц$ (oUk) до 2250 $гц$ (ik , ig). Авторы [184] считают желательным иметь значения $F_{2н}$ и $F_{2к}$ для каждого сочетания согласного с гласным в целях использования при автоматическом опознавании звуков речи.

В работе [220] исследованы сочетания 15 согласных с 8 гласными на материале 120 симметричных слогов, произнесенных тремя лицами. Формантные частоты вычислялись каждые 8,3 $мсек$ с помощью ЦВМ. Границы между согласными и гласными определялись осмотром спектрограмм по началу и окончанию звонкости в случае глухих согласных и по разрывам в спектрограммах в случае звонких согласных. По приведенным данным и графикам диапазон значений $F_{2н}$ для b простирается от 1010 $гц$ в сочетании bU до 1900 $гц$ в bi , для d — от 1700 $гц$ (da) до 1930 $гц$ (di), для g — от 1270 $гц$ (gu) до 2270 $гц$ (gi). Колебания $F_{2к}$ более значительны, в особенности для губных (например, 770 $гц$ в ip и 2180 $гц$ в ip). Разность $|F_{2н} - F_{2к}|$ в одном и том же слове достигает 250—280 $гц$ (в слогах bib , rip). По точкам пересечений кривых, характеризующих кривизну переходного участка гласных (в сочетаниях с определенными согласными), с линией нулевой кривизны определены значения локусов. $F_{2л}$ для p , b , f , v равно 1000 $гц$, для s , z — 1500, для t , d — 1700, для $\int$, $t\int$ — 1900 и для k , g > 2200 $гц$.

Зависимость перехода гласных от качества (наименования) соседних с исследуемым слогом гласных по данным наблюдения спектрограмм типа «видимая речь» рассматривалась для английской и шведской речи Оманом [200]. Речевой материал состоял из 18

английских и 75 шведских звуко сочетаний типа гласный — звонкий взрывной — гласный, произнесенных 1—2 лицами по 5 раз. Из-за влияния различных соседних гласных диапазон значений $F_{2н}$ и $F_{2к}$ для одного слога может доходить до 400—500 гц. Например, для слога *ga* $F_{2н} = 1035$ гц в сочетании *uga* и 1600 гц в сочетании *iga*.

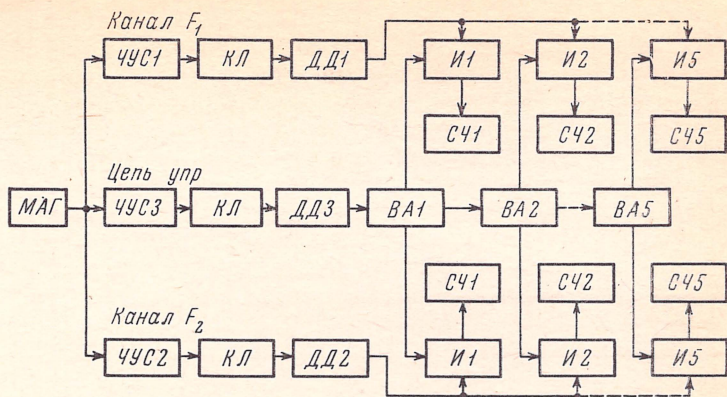
Как видно из краткого обзора, для русской речи имеются лишь неполные качественные данные о характере переходных сегментов в сочетаниях твердый согласный — гласный. Более полно исследованы слоги с мягкими согласными, но и в этом случае необходимо иметь количественные данные об изменениях формантных частот по большому числу слов, произнесенных многими лицами.

Исследования переходов английской речи проводились с участием небольшого числа дикторов, в связи с чем, как указывают авторы этих работ, полученные результаты следует считать приближенными. Тем не менее, если сравнить значения $F_{2н}$ и $F_{2к}$ в [184, 220] и сопоставить их с качественными данными о переходах в [162], то, учитывая возможные индивидуальные различия дикторов, можно заметить удовлетворительное совпадение большей части приведенных в этих работах результатов. Так, например, полученные в работах [184, 220] значения $F_{2н}$ для сочетания *gu* составляют соответственно 1325 и 1270 гц, для *pU* — 1075 и 1020 гц, для *da* — 1535 и 1700 гц, для *pI* — 1725 и 1750 гц, для *bi* — 1780 и 1900 гц.

Исследование начальных переходных сегментов гласных звуков

Задача определения параметров переходов гласных распадается на две части: нахождение временного положения начала перехода (условной границы между согласными и гласными) и определение последовательности значений параметров на протяжении перехода.

При исследовании начальных глухих взрывных в сочетаниях с гласными [96] (см. также § 6 этой главы) было установлено, что длительность шумного сегмента может служить признаком для различения *k* от *n* и *m*. Эта длительность определялась протяженностью отрезка сигнала от начала взрыва до первого импульса клиппированного сигнала длительностью $\theta \geq \theta_0$. При $\theta_0 = 0,9$ мсек лишь в четырех случаях из 160 граница сдвигалась в сторону гласного. Этот способ членения был использован затем для автоматического членения взрывных и гласных при опознавании начального *m* [99] и нашел применение в данном исследовании для членения всех исследуемых сочетаний согласных с гласными. Параметры переходных сегментов определялись по формантным частотам F_1 и F_2 . Значения F_1 и F_2 получались методом, аналогичным описанному в § 3.



Р и с. 16

Блок-схема устройства, выделяющего усредненные значения формантных частот

МАГ — магнитофон типа МЭЗ-28А; ЧУС — частотнозависимый усилитель; КЛ — клиппер; ДД — дискриминатор импульсов по длительности; ВА — выделение анализируемого участка сигнала; СЧ — счетчик; И — элемент совпадения

Речевой материал состоял из 275 слов, содержащих все используемые сочетания глухих согласных и звонких взрывных с гласными в первом ударном слоге. Слова были разбиты на 9 групп, каждую из которых произносили от 12 до 16 лиц. Кроме того, использовалось 131 слово из имевшихся других записей. Общее число дикторов — 71 (44 мужчины, 27 женщин), реализаций слов — 5140.

Блок-схема устройства, выделяющего усредненные значения формантных частот отдельных участков гласных, приведена на рис. 16. Последовательность анализируемых участков гласного начинает выделяться по прохождении первого импульса клиппированного сигнала длительностью $\theta \geq \theta_0 = 0,9$ мсек. Длительность анализируемого участка $\Delta t_1 = \Delta t_2 = 10$ мсек, $\Delta t_3 = \Delta t_4 = \Delta t_5 = 40$ мсек. Поступление импульса с $\theta \geq \theta_0$ фиксируется дискриминатором ДДЗ, посылающим сигнал в блок выделения анализируемого участка ВА1. В течение времени Δt_1 блок ВА1 подает напряжение в элементы совпадения И1 и этим включает сигналы с каналов F_1 и F_2 на счетчики СЧ1. По окончании времени Δt_1 включается блок ВА2 и т. д. По показаниям счетчиков ΔF_{1i} и ΔF_{2i} определяются усредненные значения F_{1i} и F_{2i} соответствующих участков гласного

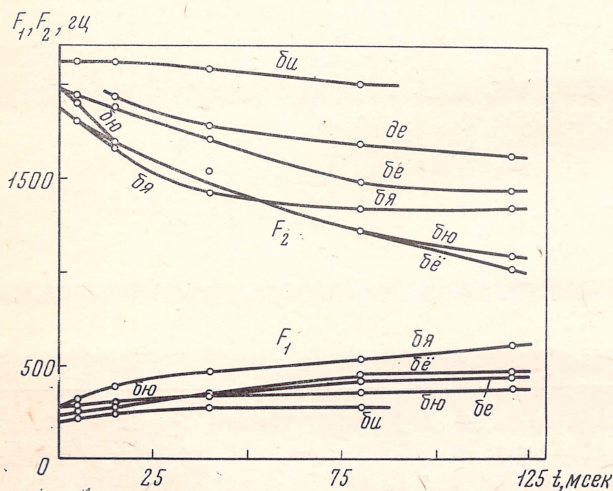
$$F_{1i} = \Delta F_{1i} / \Delta t_i, \quad F_{2i} = \Delta F_{2i} / \Delta t_i.$$

Допуская, что изменения формантных частот в начале переходного сегмента пропорциональны времени, определим начальные

и F_{11} в среднем на 50—100 гц ниже, чем после глухих согласных. Губные звуки $б, п, ф$ характеризуются двумя значениями $F_{2н}$ в диапазонах 800—900 гц (в сочетаниях с гласными заднего ряда $у, о$) и 1100—1300 гц (в сочетаниях с $а, э, ы$). Для переднеязычных $д, т, с, ш$ средние значения $F_{2н}$ занимают диапазон 1300—1400 гц при $у, о$ и близки к квазистационарным значениям F_{24} в остальных случаях. Заднеязычные $г, к, х$ характеризуются $F_{2н} \cong 800$ гц при $у, о$ и $F_{2н} \cong 1450$ гц при $а$. Сочетания $г, к, х$ с гласными $э, ы$ не исследовались, как не свойственные русскому языку.

На рис. 17—19 изображены кривые F_1 и F_2 для начальных переходных сегментов $у, а, э, ы$, расположенных после различных по месту образования согласных $б, д, г$. Отсчет времени производится от условного начала гласного. Кривые F_2 для $о$ не приведены, так как они близки соответствующим кривым для гласного $у$. Значения F_{23} близки к F_{24} , за исключением сочетаний губных согласных с $ы$, имеющих длительный переход, охватывающий всю начальную часть гласного (80—100 мсек). Длительность остальных переходных сегментов равна 40—50 мсек. На рис. 19 добавлена кривая F_2 для $тэ$, как более типичная для группы переднеязычных согласных.

В табл. 13 приведены средние значения $F_{12}, F_{15}, F_{22}, F_{25}$ для сочетаний мягких согласных с гласными. Значения $F_{1н}$ и $F_{2н}$ для «мягких» гласных (именуемых так для краткости записи) мало зависят от окружающих звуков и равны, примерно, 250 и 2000 гц. Несколько отличные значения ($F_{1н} > 250, F_{2н} < 2000$) имеют некоторые реализации 'а. В ряде случаев, в особенности при негубных согласных, F_2 запаздывает по сравнению с F_1 (позднее достигает



Р и с. 20

Т а б л и ц а 12

Средние значения формантных частот переходных и квазистационарных сегментов «твердых» гласных

Гласные F_{1i}, F_{2i}		Начальные согласные										
		б	п	ф	д	т	с	ц	ш	з	к	х
у	F_{1H}	285	350	390	315	380	385	305	370	315	370	390
	F_{11}	290	335	370	310	370	370	315	360	310	350	375
	F_{14}	325	340	360	320	360	335	335	335	325	335	330
	F_{2H}	910	900	990	1400	1290	1360	1330	1320	780	800	820
	F_{21}	880	880	940	1320	1240	1300	1270	1240	770	780	790
	F_{24}	770	820	880	980	810	890	860	860	800	770	860
о	F_{1H}	325	410	395	410	440	430	465	420	345	410	375
	F_{11}	355	420	405	410	440	440	470	450	350	420	390
	F_{14}	505	480	510	505	515	525	540	510	500	490	500
	F_{2H}	860	780	800	1410	1320	1330	1370	1260	760	780	820
	F_{21}	810	750	810	1330	1260	1270	1290	1210	760	790	830
	F_{24}	790	760	860	810	870	910	960	950	790	880	880
а	F_{1H}	520	650	580	490	570	490	530	490	390	510	590
	F_{11}	570	680	630	510	595	520	530	530	420	555	640
	F_{14}	920	950	970	820	880	890	970	880	830	880	880
	F_{2H}	1100	1080	1160	1540	1460	1430	1420	1520	1490	1450	1390
	F_{21}	1130	1130	1170	1510	1460	1420	1420	1490	1460	1430	1370
	F_{24}	1290	1340	1290	1410	1430	1370	1370	1390	1410	1390	1350
э	F_{1H}	445	520	500	300	430	420	440	325			
	F_{11}	450	515	510	345	440	430	460	370			
	F_{14}	495	540	530	465	490	495	505	480			
	F_{2H}	1270	1250	1360	1830	1650	1460	1500	1600			
	F_{21}	1280	1280	1360	1750	1590	1480	1510	1570			
	F_{24}	1450	1420	1440	1680	1510	1520	1480	1530			
е	F_{1H}	305	390	340	260	305	340	390	275			
	F_{11}	305	385	340	290	310	320	370	300			
	F_{14}	290	330	365	295	310	300	340	315			
	F_{2H}	1130	1130	1230	1930	1790	1630	1620	1770			
	F_{21}	1160	1170	1240	1930	1810	1650	1640	1750			
	F_{24}	1680	1710	1580	1990	1920	1880	1810	1900			

порогового значения). В этих случаях значения F_{21} получаются заниженными, так как начало отсчета связано с появлением F_1 . Поэтому переходные сегменты представлены в табл. 13 величинами F_{12} и F_{22} . Значения F_{15} и F_{25} для 'э в значительной мере зависят от твердости или мягкости последующих согласных. В среднем при

Таблица 13

Средние значения формантных частот начальных сегментов «мягких» гласных

Гласные F_{1i}, F_{2i}	Начальные согласные									
	б'	п'	ф'	д' дз'	т' ц'	с'	ц	г'	к'	х'
у	F_{12}	310		290	285	305	260	355		
	F_{15}	380		370	315	310	320	340		
	F_{22}	1700		1730	1760	1630	1730	1400		
	F_{25}	1080		1020	1090	1000	970	1100		
о	F_{12}	270	340	355	320	340	345	335		
	F_{15}	470	510	500	480	480	480	490		
	F_{22}	1690	1720	1540	1680	1540	1730	1410		
	F_{25}	1020	1050	1060	1020	980	1040	960		
а	F_{12}	390	370		395	440	445	490		
	F_{15}	605	590		550	645	570	690		
	F_{22}	1660	1700		1770	1520	1500	1540		
	F_{25}	1340	1350		1290	1380	1320	1420		
э	F_{12}	285	280	310	305	300	340	270	290	330
	F_{15}	450	440	445	365	430	435	420	420	420
	F_{22}	1870	1860	1870	1940	1840	1930	1880	1870	1780
	F_{25}	1440	1500	1540	1610	1510	1510	1740	1590	1560
и	F_{12}	255	265	260	250	260	255	260	255	270
	F_{14}	270	290	285	275	235	295	290	280	300
	F_{22}	2120	1970	2020	2080	1890	1940	2090	2030	2020
	F_{24}	2000	2030	1930	2030	2070	1850	1990	2010	1950

последующих твердых согласных $F_{15} = 440$, $F_{25} = 1490$ гц, при мягких $F_{15} = 370$, $F_{25} = 1720$ гц. Большой разброс этих значений для 'э вызван главным образом тем, что большинство сочетаний б'э оканчивались твердыми согласными, а д'э и х'э — мягкими.

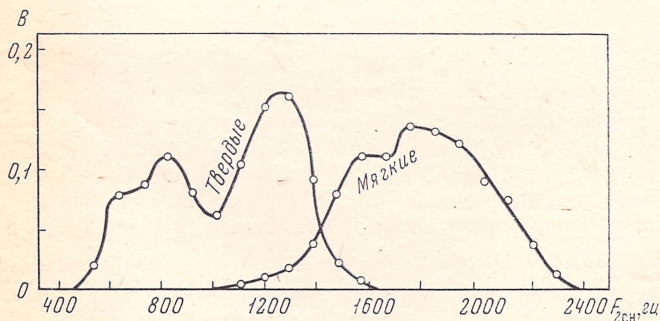
Кривые F_1 и F_2 начальных сегментов гласных, расположенных после б', а также F_2 для сочетаний д'э приведены на рис. 20. Кривые F_2 для б'э и д'э характеризуют соответственно случаи, когда после 'э следуют твердые или мягкие согласные.

**Признаки согласных звуков,
вытекающие из переходных сегментов в гласных**

Рассмотрим параметры переходных сегментов гласных, пригодные к использованию в качестве признаков твердости или смягчения, и места образования предыдущих согласных. Весьма важной для

русской речи является информация о наличии смягчения согласных. Эта информация, как известно, содержится как в некоторых самих согласных, так и в *и*-образном переходном сегменте последующих гласных. Однако, как следует из приведенных выше данных, для многих твердых согласных (в особенности язычных в сочетании с *а*, *э*, *ы*) переходный сегмент весьма близок к *и*-образному, характеризуемому низким F_1 и высоким F_2 . С другой стороны, некоторые реализации *'а* имеют *э*-образное начало. Большую часть указанных реализаций (главным образом *а* и часть *'а*), для которых при классификации по наличию *и*-образного начала будет получено ошибочное решение, можно заранее выделить по $F_{14} \geq F_{1a}$, где F_{1a} — нижняя граница значений F_1 для гласного *а*.

Начальный *и*-образный сегмент может быть охарактеризован длительностью анализируемого участка, считая от начала гласного — Δt_n , и средними значениями формантных частот на этом участке — $F_{1сн}$ и $F_{2сн}$, а также значением F_{13} . При выборе Δt_n необходимо учесть, что при слишком коротком участке некоторые сочетания переднеязычных *д*, *т*, *с*, *ц* с гласными *у*, *о* могут классифицироваться как мягкие. Кроме того, из-за некоторого «запаздывания» F_2 по отношению к F_1 в отдельных реализациях *и* значения F_{21} для этих реализаций могут получаться заниженными. При слишком большом Δt_n значения $F_{1сн}$ и $F_{2сн}$ для неоднородных мягких гласных сближаются и участок начинает терять *и*-образный характер, а сочетания *б*, *п*, *ф* с *э* и *ы* начинают классифицироваться как «мягкие». Исходя из указанных данных было определено значение $\Delta t_n = 20$ мсек, откуда $F_{1сн} = 0,5(F_{11} + F_{12})$, $F_{2сн} = 0,5(F_{21} + F_{22})$. Из числа этих параметров намного более информативным является $F_{2сн}$. На рис. 21 приведены нормированные кривые плотности распределения (частоты) $B(F_{2сн})$ для твердых и мягких гласных, у которых $F_{14} < F_{1a} = 675$ гц. На оси ординат отложено отношение числа реализаций, имеющих значения $F_{2сн} \pm 50$, к общему числу реализаций соответственно твердых или мягких звуков. Гласные *э* и *ы* после переднеязычных согласных имеют высокие значения $F_{2сн}$ и поэтому, как исключение из общего правила, включены



Р и с. 21

в состав мягких гласных. Наличие двух максимумов в одной из кривых $B(F_{2\text{сн}})$ вызвано значительно меньшими значениями $F_{2\text{сн}}$ для y, o после губных b, n, ϕ и заднеязычных g, k, x по сравнению с величинами $F_{2\text{сн}}$ для остальных гласных.

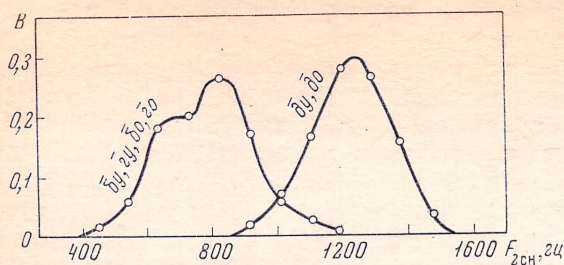
Как видно из кривых $B(F_{2\text{сн}})$, твердые и мягкие согласные хорошо разделяются по признаку $F_{2\text{сн}}$ последующего гласного (при $F_{14} < F_{1a}$). Реализации гласных с $F_{14} \geq F_{1a}$ могут быть разделены на твердые и мягкие по признаку F_{13} , так как для a значение F_{13} больше (ввиду быстрого нарастания F_1), чем для $'a$. По описанному алгоритму было классифицировано 4802 реализаций гласных — 2997 твердых, 1805 мягких — при пороговых значениях признаков $F_{1a} = 675 \text{ гц}$, $F_{2\text{сн}} = 1430 \text{ гц}$, $F_{13} = 560 \text{ гц}$. Число ошибок составило 269, или 5,6%. Гласные ε после переднеязычных согласных выделяются из группы мягких по $F_{1\text{сн}} \geq 400$, $F_{24} - F_{14} \geq 850 \text{ гц}$. Гласные y в аналогичной позиции не могут быть достаточно хорошо отделены от мягких гласных (точнее, от u) по признакам, используемым в данной работе.

Перейдем к установлению признаков для определения места образования твердых согласных. Такими признаками в первую очередь могут быть параметры $F_{2\text{н}}$, F_{14} , F_{24} соседних гласных. Однако $F_{2\text{н}}$ необходимо вычислять. Поэтому целесообразно заменить их на значения F_{21} или $F_{2\text{сн}}$. Преимуществом $F_{2\text{сн}}$ является меньший разброс относительно средних значений и меньшая частота отсчета. По влиянию гласных на значения $F_{2\text{н}}$ их можно разделить на три категории: y, o ; a ; ε, y . Для сочетаний согласных, принадлежащих к одной группе по месту образования, и гласных каждой из этих категорий значения $F_{2\text{н}}$ близки друг к другу.

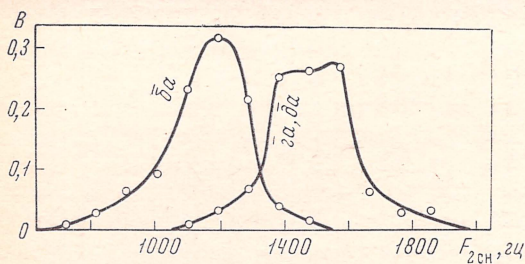
На рис. 22 приведены кривые плотности распределения $B(F_{2\text{сн}})$ для гласных y, o в зависимости от места образования предыдущих согласных. Хотя средние значения $F_{2\text{сн}}$ для губных и заднеязычных несколько различаются, распределения $F_{2\text{сн}}$ обеих групп в значительной мере пересекаются и представлены одной кривой (обозначенной взрывными $\bar{b}$, $\bar{g}$). Надежность разделения переднеязычных от остальных согласных на материале 969 реализаций составила 94,7%.

На рис. 23 приведены кривые $B(F_{2\text{сн}})$ для гласных a . Губные согласные b, n, ϕ различаются от язычных с надежностью 91,4% на материале в 439 реализаций. Наличие двух максимумов $B(F_{2\text{сн}})$ в правой кривой объясняется сдвигом распределений переднеязычных и заднеязычных согласных. Кривые $F_{2\text{сн}}$ для гласных ε и y даны на рис. 24. Надежность разделения губных и язычных согласных на материале 898 реализаций составила 93,9%.

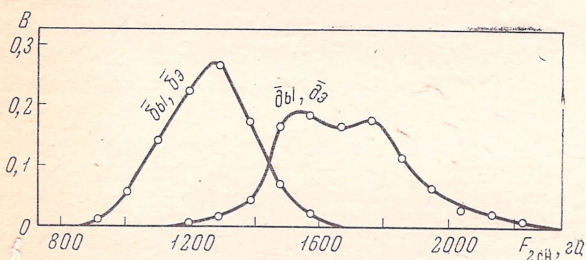
Использование признака средней частоты начального участка $F_{2\text{сн}}$ позволяет разделить твердые согласные на две группы (но не на три основные группы) по месту их образования. Надежность разделения может быть повышена, если исходить из распределений $F_{2\text{сн}}$ отдельно для каждого гласного и отдельно для звонких и глу-



Р и с. 22



Р и с. 23

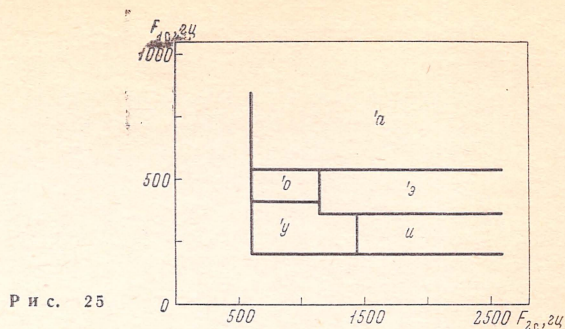


Р и с. 24

хих согласных. Из табл. 13 следует, что влияние места образования мягких согласных на переходной сегмент соседнего гласного мало. Поэтому в переходном сегменте этих гласных не содержится достаточной информации для определения места образования мягкого согласного.

Опознавание гласных, расположенных после мягких согласных

Полученные значения F_{1i} и F_{2i} дают возможность осуществить опознавание мягких гласных (в исследованных сочетаниях) с большей надежностью, чем по последовательности отрезков, не связанной жестко с началом гласного. В то время как начальный участок гласного свидетельствует главным образом о смягчении предыдущего согласного и звуко сочетания в целом, конечный участок переходного сегмента вместе с квазистационарной частью со-



Р и с. 25

держит основную информацию о принадлежности (тембре) гласного. На рис. 25 приведены области распределения мягких гласных на плоскости $F_{1c}F_{2c}$, где F_{1c} и F_{2c} — средние значения формантных частот за время $\Delta t_4 + \Delta t_5$. Границы областей проведены исходя из распределений F_{1c} и F_{2c} для мягких гласных и приблизительно соответствуют границам для твердых гласных, представленным на рис. 6. Надежность опознавания мягких гласных по признакам F_{1c} и F_{2c} приведена в табл. 14.

Т а б л и ц а 14

Надежность опознавания «мягких» гласных в процентах

Произнесено	Опознано				
	'у	'о	а	'э	и
'у	84,5	3,3	—	10,5	1,7
'о	4,9	75,5	7	12,6	—
'а	—	2,6	85,4	12	—
'э	1,7	2	0,7	88,4	7,2
и	3,3	—	—	8,4	88,3

Влияние смягчения последующего согласного мало сказывается на 'у, 'о, и и довольно значительно на 'а и 'э. В этих случаях 'а приближается к 'э, а 'э — к и. Варианты 'а и 'э перед мягкими согласными не образуют отдельных областей на плоскости $F_{1c}F_{2c}$, мало пересекающихся с другими областями. Значительное большинство реализаций 'а и 'э в этих случаях попадает в свои области, остальные реализации переходят в соседние области с меньшим F_{1c} . Надежность опознавания мягких гласных с учетом частоты встречаемости составляет 87,7%, так как подавляющую часть этих гласных (около 90%) составляют 'э и и, а 'у и 'о встречаются значительно реже.

Для определения значений $F_{2л}$ были построены кривые F_2 , усредненные для сочетаний согласных, одинаковых по месту образования, с каждым гласным. На рис. 26 и 27 приведены указанные кривые для сочетаний губных $\bar{b}$, $\bar{n}$, $\bar{f}$ и переднеязычных $\bar{d}$, $\bar{t}$, $\bar{s}$, $\bar{c}$ со всеми гласными. Точки (точнее, центры областей) пересечения продолжений переходных сегментов F_2 соответствуют локусам $F_{2л}$. Полученное значение $F_{2л} = 1740$ гц для переднеязычных согласных хорошо согласуется с данными, приведенными в [133, 220]. Локус для губных получился равным 1020 гц, что совпадает с данными для естественной речи [220] и не совпадает со значением локуса для синтетических слогов, равным 700 гц [133]. Положение локуса (или локусов) $F_{2л}$ для заднеязычных $\bar{g}$, $\bar{k}$, $\bar{x}$ в данном исследовании не обнаружено. Это объясняется известной зависимостью места образования заднеязычных от типа последующего гласного (в особенности от того, является ли этот гласный передним или задним). Кроме того, сочетания заднеязычных согласных с гласными были представлены лишь четырьмя кривыми F_2 , что также затруднило определение значений локусов в этом случае.

Результаты исследования переходных сегментов гласных, проведенного автором совместно с А. А. Григоряном, использованы при опознавании гласных и взрывных по последовательности сегментов (см. § 7 этой главы).

Рис. 27

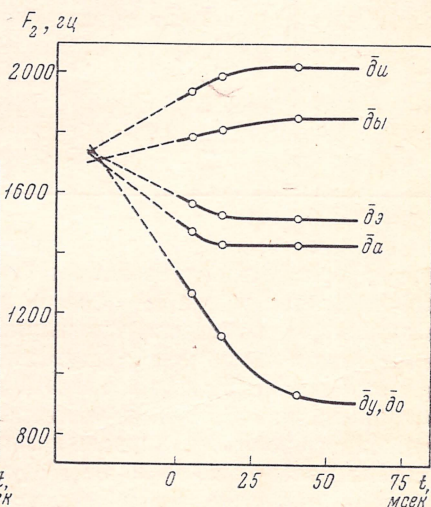
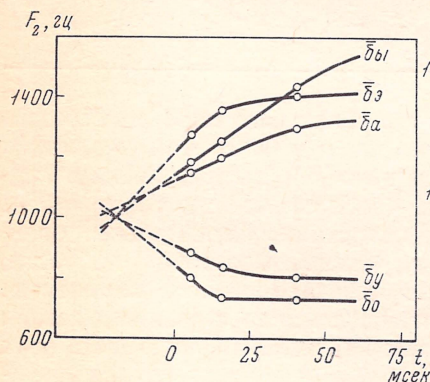


Рис. 26



Как известно, большая часть речевой информации содержится в согласных звуках, большинство которых обладает шумовой составляющей. В связи с этим представляет значительный интерес исследование шумов, характеризующих отдельные звуки, с целью использовать полученные данные для опознавания речи. Различия в свойствах шумовой составляющей в зависимости от места образования звука должны наиболее четко выявляться у глухих щелевых *с, ш, ф, х*, обладающих сравнительно длительным квазистационарным сегментом.

Нахождение признаков глухих щелевых даст также возможность приблизиться к опознаванию остальных шумных, ввиду близости характера шумов звуков, различающихся по способу образования, но одинаковых по месту образования (например, *с, ц, т, з, д*). Эта близость сказывается на восприятии шумных звуков человеком. Так, по данным Дергача, Дукельского [32, 35], при укорочении щелевых в слогах типа СГ (согласный — гласный) до величины, когда длительность шумной части начинает соответствовать взрывным звукам, *ф* воспринимается как *п*, *с* и *ш* — как *т, х* (при большой длительности) — как *к*. Аналогичный эффект имеет место и при укорочении мягких согласных *с', ф', х'*. Опознавание человеком щелевых (в слогах типа СГ) происходит не только по шумовой части, но и частично по переходному участку, включающему начало соседнего гласного. Наличие этого участка необходимо при восприятии звука *ф'*.

Для опознавания глухих щелевых необходимо использовать как групповые признаки — отсутствие звонкости, шумность (большая плотность нулей), малая крутизна нарастания сигнала, большая длительность, так и признаки, вытекающие из свойств шума и позволяющие различать звуки внутри группы. Представляет интерес исследовать шум щелевых по клиппированному речевому сигналу, в котором признаки звуков могут содержаться только в тех или иных свойствах расположения нулей. Как следует из данных, приведенных в табл. 8, 9 (гл. 2), при полосе пропускания 0—7 кГц и некотором ослаблении нижних частот разборчивость большинства щелевых клиппированной речи находится в пределах 80—100% и лишь для начальных *с* и *х'* разборчивость уменьшается до 60—70%. Неправильно опознанные слушателями звуки воспринимались: *ф* как *ш* и *х*, *х'* как *ф'*, *ф'* как *х'* и *ц*, начальные *с* как *ц*. В последних случаях очевидно сказывалась замена при клиппировании пологого нарастания сигнала — одного из групповых признаков щелевых — на резкое нарастание, свойственное аффрикатам *ц* и *ч*. Все же разборчивость щелевых в целом остается после клиппирования достаточно хорошей.

Большинство работ по исследованию щелевых (как, впрочем, и других звуков речи) основано на спектральном анализе речевого сигнала. Тарноци [225], изучавший спектры восьми щелевых, встречающихся в западноевропейских языках, на материале от четырех дикторов, отметил, что должны приниматься во внимание различия не только по спектру, но и по интенсивности. Так, например, f отличается от s и $\int$ меньшей интенсивностью, в среднем на 12,5 дБ. Спектр щелевых зависит также от тональности соседних гласных. При сравнении $\int$, расположенного между низкими и высокими гласными, спектр в первом случае несколько смещен вниз. Варшавский и Литвак [10] определяли диапазоны частот, характерные для s , $ш$, ϕ , x . Результаты прослушивания 15 лицами «белого» шума, ограниченного фильтрами нижних и верхних частот, показали, что $ш$ характеризуется полосой 1,2—6,3 кГц, s — 4,2—8,6 кГц, ϕ — 7—12 кГц, x — 0,4—1,2 кГц.

В работе Гейнца и Стивенса [168] определены значения полюсов P_1 , P_2 и нулей Q_1 передаточных функций цепей для синтеза щелевых. Для $\int$ — $P_1 = 2200$ — 2400 гц (в зависимости от контекста) при ширине 400—600 гц, $P_2 = 4300$ — 5400 (ширина 900—1100), $Q_1 = 3400$ — 4400 (1400—1800). Для s — $P_1 = 3500$ — 6400 (600—900), $P_2 = 8000$ — 8400 (600—1400), $Q_1 = 2300$ — 3300 (1000—1300). Для f $P_1 = 6800$ — 8400 (900—1000), $P_2 = 8200$ — 12200 (900—1100), $Q_1 = 4600$ — 6800 гц при ширине 600—1100 гц. Хотя модель для синтеза щелевых весьма идеализирована, полученные на ней спектры хорошо совпадали со спектрами естественных s , $\int$, f , измеренными в [171].

Метод опознавания s , $\int$ и f по отношениям энергии в трех парах частотных полос описан Хьюзом и Халле [171]. Вначале по отношению энергии в полосах 720 гц — 10 кГц и 4,2—10 кГц щелевые делились на 2 группы: s , f и $\int$. Разделение s и f в первой группе происходило по отношению энергии в полосах 720—6500 гц и 720—2150 гц. Разделение во второй группе осуществлялось по отношению энергии в полосе шириной 500 гц, охватывающей максимум интенсивности в области 1,5—4 кГц, и в полосе 720—1370 гц. Надежность различения s , $\int$ и f на материале 190 реализаций от пяти дикторов составила около 90%. При подавлении в речевом сигнале нижних частот до 700 гц звонкие щелевые были близки соответствующим глухим.

Опознавание начальных щелевых s , $\int$, f , h , v , z с помощью так называемой нейронной логики исследовалось Нельсоном и др. [196]. Речевой сигнал разделялся на 19 каналов с помощью полосовых фильтров с низкой добротностью. При сравнении энергии в соседних каналах выделялись следующие признаки: локальные максимумы и минимумы, положительные и отрицательные наклоны

огibaющей спектра (области возрастания и убывания энергии в спектре), интенсивность энергии в каналах e_i . Вначале разделялись глухие и звонкие щелевые. В группе глухих s и f различались по положительным наклонам, наличию энергии в первых девяти каналах и разности положительных наклонов в диапазонах выше и ниже 2500 $гц$. Признаками $\int$ служат: непрерывный положительный наклон в первых 11 каналах, отсутствие подобного наклона в ВЧ-каналах, отсутствие максимума выше 4 $кгц$, большая длительность (по сравнению с h). Энергия h сосредоточена в значительной мере в области F_2 или F_1 и F_2 последующего гласного, ввиду чего реализации h были разделены на 4 группы. Каждая из этих групп имела отдельные признаки. Разделение z и v осуществлялось по распределению энергии в отдельных каналах и областям положительного наклона. Надежность опознавания начальных щелевых в сочетаниях с 10 различными гласными в слогах типа СГС, оканчивающихся согласным d , при участии 6 дикторов составила 86% для v и от 93,5 до 99,2% для остальных щелевых.

Анализ z и $\int$ по клиппированному речевому сигналу производили Сакаи и Иноуэ [210]. Ими приведены кривые плотности распределения длительностей интервалов между нулевыми пересечениями (для одного диктора). Максимум кривой для $\int$ соответствует интервалу около 0,12 $мсек$, а для z — интервалу длительностью менее 0,1 $мсек$. В работе Дошита [138] анализировались z , $\int$, h и ts , $t\int$ в открытых слогах. Интервалы между нулями классифицировались по частоте (длительности) на 14—16 каналов. Из приведенных кривых для одного диктора следует, что для z максимум распределения нулевых пересечений расположен в диапазоне 5—10 $кгц$ и мало зависит от последующего гласного, исключая заметное смещение максимума до ~ 5 $кгц$ в сочетании $зи$. Весьма значительно влияние гласного на распределение нулей для h . Максимум распределения смещается от ~ 3 $кгц$ при he до 1200 $гц$ при hi . Вайрен и Стабс [229] различали s и $\int$ по усредненной плотности нулей ρ_+ с проверкой на небольшом числе реализаций. Пороговое значение ρ_+ составляло около 5 $кгц$.

Данные по опознаванию щелевых, приведенные в опубликованных работах, основаны на анализе сравнительно небольшого числа реализаций звуков, расположенных в слогах (словах) типа СГ и СГС. Для получения более достоверных данных необходимо при анализе и проверке информативности признаков применять более разнообразный речевой материал, включающий многосложные слова и все типы встречающихся сочетаний щелевых с другими звуками (в том числе в неударных слогах), используя несколько десятков лиц в качестве дикторов.

На основании предварительного исследования было установлено, что характер шума щелевых может быть выражен в общих чертах двумя свойствами: средней длительностью интервалов между нулями, определяемой по количеству нулей за некоторое время, и диффузностью, выражающей степень разброса длительностей интервалов между нулями. Диффузность шума может определяться нормированной дисперсией длительностей интервалов. Однако для получения более простого технического решения целесообразно выразить степень разброса числом интервалов, длительность которых значительно больше и (или) значительно меньше средней длительности. Диффузность шума будет характеризоваться в этом случае лишь крайними участками кривой нормированной дисперсии. В качестве признаков шумных звуков были выбраны следующие параметры:

$$2N = 2\Delta N / \Delta t,$$

где $2\Delta N$ — число нулей на анализируемом участке длительностью Δt ;

$$n_i = \Delta n_i / \Delta t,$$

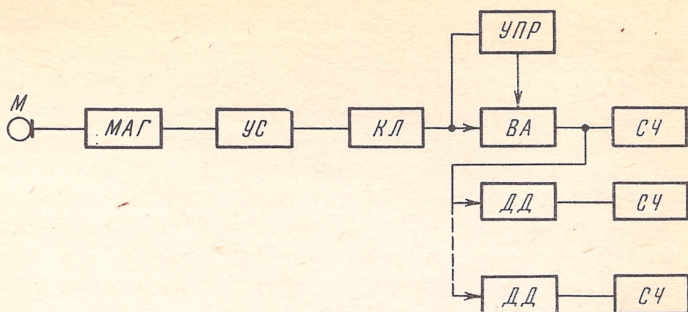
где Δn_i — число интервалов между нулями (соответствующих положительным импульсам сигнала) длительностью более $1/2f_i$ на том же участке;

$$m_i = \Delta m_i / \Delta t,$$

где Δm_i — число интервалов между нулями длительностью менее $1/2f_i$ на том же участке;

$$r_1 = n_1 + m_1; \quad r_2 = n_3 + m_2.$$

Указанные признаки выделялись из слов с помощью специального устройства, скелетная схема которого приведена на рис. 28. Выделение анализируемого участка исследуемого звука происходило с помощью схемы управления, которая начинала действовать при достижении усредненной плотностью нулей речевого сигнала одного из двух пороговых значений, характерных для щелевых. Время усреднения составляло 25 мсек. Чтобы признаки возможно меньше зависели от напряжения речевого сигнала, включение схемы управления обуславливалось также наличием достаточно большого напряжения. Участок сигнала для анализа признаков продолжительностью 40 мсек начинал выделяться спустя примерно 60 мсек после начала исследуемого звука. f_1 , f_2 и f_3 составляли соответственно 1200, 2400 и 4800 гц, $f'_1 = 9$ кгц, $f'_2 = 12$ кгц. Выделенный участок сигнала поступал на дискримина-



Р и с. 28.

Скелетная схема устройства для выделения признаков шумных звуков

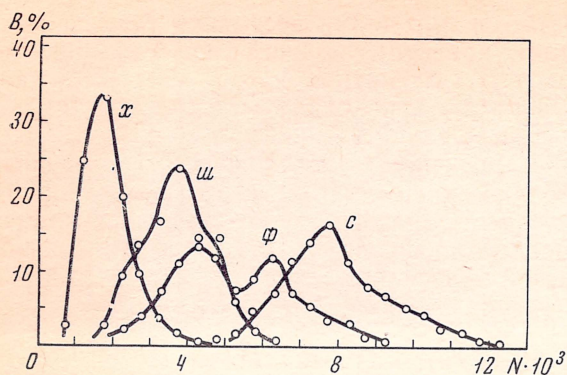
М — микрофон МД-45; МАГ — магнитофон МЭЗ-28А; УС — усилитель; КЛ — клиппер; УПР — блок управления; ВА — выделение анализируемого участка сигнала; ДД — дискриминатор; СЧ — счетчик

торы по длительности и счетчики, на которых фиксировались значения исследуемых признаков. При наличии в слове двух или более звуков с большой плотностью нулей устройство выделяло признаки первого из них.

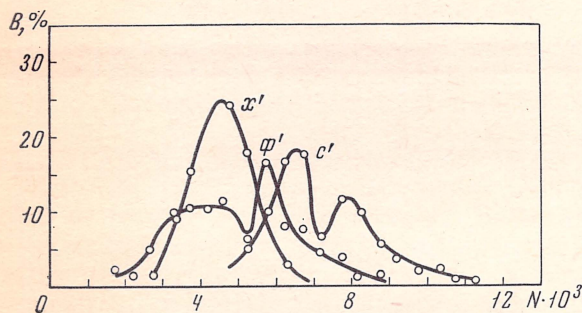
В эксперименте приняли участие в качестве дикторов 46 лиц (27 мужчин, 19 женщин) в возрасте от 17 до 53 лет, которые не проходили какой-либо подготовки. Речевой материал состоял из 93 слов, в том числе 44 многосложных. Исследуемые щелевые располагались в начале, середине и конце слов, до и после почти всех гласных. В 14 словах щелевые сочетались с другими согласными. Лишь звук *х* был представлен большей частью в начальном положении ввиду малой плотности нулей соответствующего ему сигнала. Семь лиц произносили все слова, остальные — лишь ту или иную половину общего числа слов, причем в каждой половине были сохранены основные варианты всех исследуемых звуков. У ряда лиц, говоривших сравнительно тихо, устройство не анализировало в некоторых словах звуки *ф*. Эти слова требовалось повторить более громко.

Из общего количества 2370 произнесенных слов было анализировано 2294 звука, в том числе 472с, 180с', 423ш, 358ф, 211ф', 410х и 189х'. В 51 случае устройство анализировало другие звуки, главным образом *и*, *э*, *в*, расположенные до исследуемых щелевых. В 76 случаях схема управления устройством не включалась (при слишком тихом произнесении или при величине усредненной плотности нулей ниже пороговой).

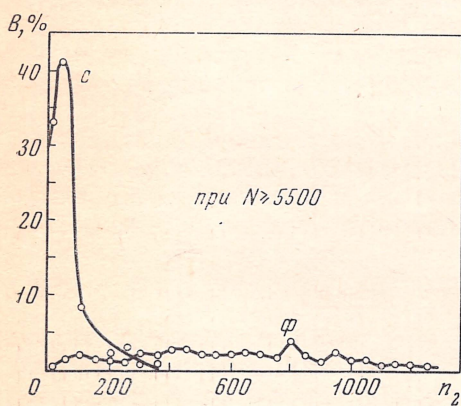
На рис. 29, 30 приведены кривые плотности распределения (частоты) числа нулей $B(N)$ исследуемых звуков, отдельно для твердых и мягких согласных. На оси ординат отложено отношение числа реализаций, имеющих значения от $N - 250$ до $N + 250$, к общему числу реализаций данного звука. По кривым $B(N)$



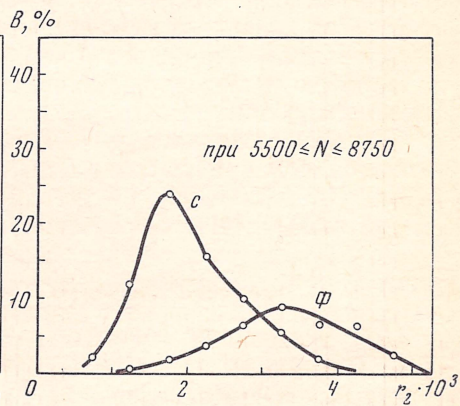
Р и с. 29



Р и с. 30



Р и с. 31



Р и с. 32

можно хорошо различать c и $ш$, c и $х$ и несколько хуже ϕ и $х$, $х$ и $х'$. Пользуясь этими кривыми, можно всю совокупность реализаций твердых щелевых разделить на четыре области, включающие: при $N < 1750$ — $х$; при $1750 \leq N < 2750$ — $ш$, $х$, при $2750 \leq N < 5500$ — $ш$, ϕ , при $N \geq 5500$ — c , ϕ .

Указанные пороговые значения N могут быть использованы также для различения мягких щелевых (с несколько худшими результатами), причем область значений $х'$ в значительной мере совпадает с областью для $ш$. Для звуков c' и ϕ' кривые охватывают примерно тот же диапазон значений N , что и для соответствующих твердых фонем.

Дальнейшее различение отдельных щелевых может быть основано на диффузности шума этих звуков. На рис. 31, 32 приведены кривые $B(n_2)$ и $B(r_2)$ для c и ϕ при $N \geq 5500$.

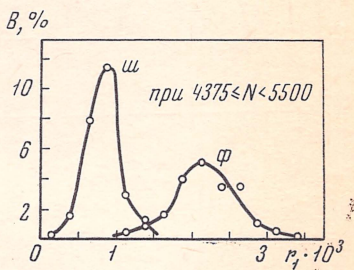
Признак n_2 хорошо отражает меньшую диффузность шума c по сравнению с ϕ и позволяет разделить реализации, имеющие $N \geq 5500$, на области c и ϕ . Некоторая часть ϕ , имеющих значение n_2 ниже пороговой величины и попадающих поэтому в область c , может быть выделена из этой области по признаку r_2 .

Для разделения $ш$ и $х$ (при $1750 \leq N < 2750$), а также $ш$ и ϕ (при $2750 \leq N < 5500$) следует использовать признак r_1 , указывающий на меньшую диффузность шума $ш$ по сравнению с $х$ и ϕ . С увеличением N увеличивается также значение r_1 как для $ш$, так и для $х$ и ϕ , ввиду чего одним пороговым значением этого признака не удастся различить рассматриваемые звуки. Поэтому следует разделить диапазон N на несколько частей и в каждой из них выбрать пороговую величину r_1 . На рис. 33 приведены для примера кривые плотности распределения $B(r_1)$ для $ш$ и ϕ при $4375 \leq N < 5500$.

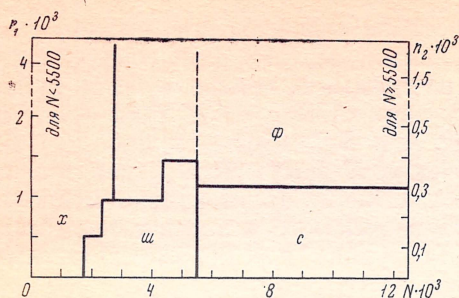
Различение мягких щелевых c' , ϕ' и $х'$ может быть произведено почти аналогичным образом. Признак N разделяет реализации мягких звуков на две группы: c' и ϕ' , ϕ' и $х'$. Различение c' и ϕ' производится по n_2 и r_2 , различение ϕ' и $х'$ — по r_1 и n_2 .

Надежность опознавания глухих щелевых

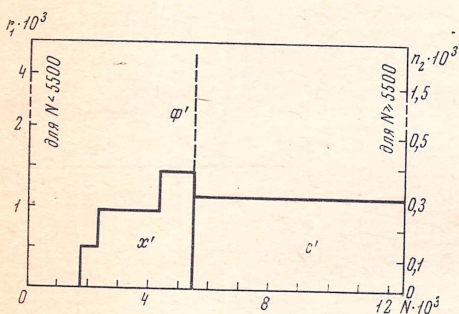
При разделении совокупности реализаций исследуемых звуков на области по полученным признакам следует учитывать частоты встречаемости этих звуков. По данным Ёлкиной и Юдиной [36], из числа русских щелевых наибольшую частоту встречаемости имеет c (3,27%), а наименьшую — ϕ' и $х'$ (оба по 0,1%). Остальные



Р и с. 33



Р и с. 34



Р и с. 35

глухие щелевые имеют частоту встречаемости в пределах 0,92—1,79%.

На плоскостях признаков Nr_1 и Nn_2 (по данным о плотности распределения этих признаков) было произведено разделение реализаций твердых щелевых на области $с$, ϕ , $ш$, $х$ (рис. 34) и мягких щелевых — на области $с'$, ϕ' и $х'$ (рис. 35). На обоих графиках оказалось возможным использовать одни и те же пороговые значения признаков. В дополнение к графику, приведенному на рис. 35, реализации, имеющие $r_2 < 3625$ и оказавшиеся в области $с'$, выделялись в область ϕ' . Последняя была также увеличена за счет включения в нее реализаций с $n_2 \geq 950$ из области $х'$. Надежность опознавания и характер ошибок, получаемых при разделении реализаций щелевых на указанные области, приведены в табл. 15 и 16.

Как следует из таблиц, с меньшей надежностью различаются редко встречаемые ϕ' и $х'$. Средняя надежность различения глухих щелевых с учетом частоты встречаемости составляет около 91%.

Надежность различения зависит как от произношения отдельных лиц, так и от типа слов. У диктора С. все ϕ отождествлялись с $ш$. На 8 «худших» дикторов приходится 15,1 всех реализаций и 27% всех ошибок. Результаты большинства из них могут быть улучшены при более четком и внятном произношении, достигаемом небольшой тренировкой. Почти половина ошибок при выделе-

Таблица 15

Надежность опознавания твердых щелевых в процентах

Произнесено	Опознано			
	<i>с</i>	<i>ф</i>	<i>ш</i>	<i>х</i>
<i>с</i>	94,1	4,9	0,6	0,4
<i>ф</i>	5,9	85,8	3,6	4,7
<i>ш</i>	1,6	5,0	90,3	3,1
<i>х</i>	0	8,0	4,9	87,1

Таблица 16

Надежность опознавания мягких щелевых в процентах

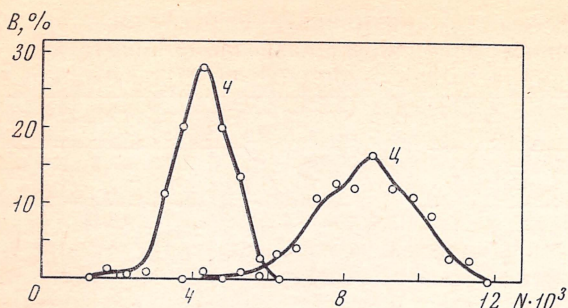
Произнесено	Опознано		
	<i>с'</i>	<i>ф'</i>	<i>х'</i>
<i>с'</i>	90	3,3	6,7
<i>ф'</i>	9,0	81,5	9,5
<i>х'</i>	10,6	7,4	82

лении *с* (13 из 28) падает на слово *ласточка*. Это объяснилось уменьшением длительности *с* перед взрывными (особенно заметным в данной ситуации), в связи с чем для анализа выделялся не квазистационарный, а конечный переходный участок звука. В другом аналогичном звукосоочетании (слове *проскок*) уменьшение длительности *с* приводило лишь к сдвигу реализации на плоскости Nn_2 в направлении границы области *с* без выхода за ее пределы. Исключение слова *ласточка* повышает надежность различения *с* для оставшихся 18 слов с 94,1 до 96,7%. Звук *ф* хуже всего различается в слове *Мефистофель* — 40% реализаций попадали в область *х'*. На 16 «худших» слов (большой частью многосложных) приходится 17,5% всех реализаций и 40,5% всех ошибок.

В данном исследовании опознавание твердых и мягких щелевых проводилось отдельно для каждой из этих групп звуков. Признаками для различения обеих групп служат отсутствие или наличие *и*-образного переходного сегмента в соседнем гласном и характер шума [35]. Первый из них играет решающую роль при опознавании *ф*. В характере шума этого звука нет, по-видимому, признаков мягкости. Что касается *с* и *х*, то здесь важную роль играет также характер шума. Исследуемые признаки выявили различия для пары *х* и *х'*, но не выявили их для *с* и *с'*. Совпадение границ областей твердых и мягких щелевых может быть использовано для различения этих звуков независимо от степени мягкости при опознании сравнительно большого набора, когда наряду с групповыми признаками необходимо применять признаки, выделяющие некоторые звуки внутри групп. Следует при этом учесть, что *х'* будет совпадать не с *х*, а с *ш*.

К опознаванию аффрикат и звонких щелевых

На небольшом речевом материале было проведено предварительное исследование информативности полученных признаков для различения звонких щелевых *з*, *ж* и аффрикат *ц*, *ч*. Никаких изме-



Р и с. 36

нений параметров устройства по выделению признаков (рис. 29) при этом не производилось. В результате эксперимента оказалось, что указанные пары звуков хорошо различаются по тем же признакам, что и глухие щелевые, с сохранением границ областей на плоскостях Nr_1 и Nn_2 , приведенных на рис. 34 и 35. Реализации *ч* и *ж* попадают в область *ш*; *ц* — в область *с*; *з* — в область *ш* и *ф*. Различение звуков, входящих в разные группы (например, *ш*, *ч* и *ж*), должно производиться по групповым признакам. На рис. 36 приведены плотности распределения числа нулей $B(N)$ для звуков *ц* и *ч*, полученные по 340 реализациям от 22 дикторов. Что касается звонкого щелевого *в*, то его шумовая составляющая настолько слаба, что даже в слогах под ударением (в словах *восемь*, *Вася*) устройство реагировало на нее лишь в 8,8% случаев. Этот результат совпадает с данными Фанта, по которым удаление шумовой составляющей *в* влияет больше на натуральность этого звука, чем на разборчивость [88]. Решающую роль при опознавании *в* играет, по-видимому, голосовой источник звука. Результаты данного исследования были приведены в [103].

На основе анализа длительных шумных на протяжении звука и описанных выше признаков N , n_1 , n_2 , m_1 Рудный и Трунин-Донской разработали алгоритм опознавания *с*, *ш*, *ф*, *х*, *х'*, *з*, *ж*, *ц*, *ч*, *т'*, *к'*. Положение квазистационарного сегмента определяется в зависимости от положения исследуемой фонемы (в начале, середине или конце слова) и длительности. При принятии решения методом логического дерева учитывается наличие смычки в начале и конце фонемы, длительность шумового сегмента tr , величина и положение $N_{\max}$ (у аффрикат и взрывных $N_{\max}$ расположен в начале фонемы), разность $N_{\max} - N_1$, динамика n_1 , n_2 , m_1 . Надежность опознавания шумных звуков на материале 3244 реализаций (на которые срабатывал признак p), полученном при участии 40 лиц, составила около 90%. Изменения признака N на протяжении звука были проверены Высоцким и Кулагиным на ЦВМ БЭСМ-6. Речевой сигнал вводился через преобразователь аналог — код [17].

Исследования щелевых ограничивались до последнего времени анализом и получением признаков квазистационарных сегментов. Между тем, влияние соседних звуков, заметное на параметрах квазистационарных сегментов щелевых, тем более сказывается на переходных сегментах. Характер распределения длительностей интервалов между нулями в переходах щелевых зависит, как уже указывалось, от типа соседнего гласного и содержит некоторую информацию о принадлежности этого гласного. Использование переходов щелевых может оказаться также полезным при определении принадлежности соседних взрывных. Как отмечает Шварц [212], при удалении в словах *last*, *rasp*, *task*, конечного взрыва слушатели узнавали 75% *p*, 48% *k* и лишь 5% *t*.

§ 6. Оpoznание взрывных по признакам шумовых сегментов

Краткий обзор

Взрывной звук в сочетании с гласным характеризуется последовательностью трех основных сегментов: смычки (глухой или тональной), шумового сегмента (собственно взрывного) и переходного сегмента последующего гласного. Для конечных взрывных после перехода гласного следует смычка и шумовой сегмент. Информация о взрывном содержится также в переходных сегментах соседних щелевых. Групповыми признаками взрывных являются: смычка, крутизна нарастания сигнала, шумность, длительность шумового сегмента. Признаками отдельных взрывных, характеризующими место образования этих фонем, могут служить параметры шумового сегмента (распределение энергии по спектру или число и распределение интервалов между нулями) и параметры переходного сегмента соседнего гласного или щелевого. Наличие смычки является обязательным признаком взрывного. Шумовой сегмент или переходная область соседнего звука могут в отдельных случаях отсутствовать.

Малая интенсивность, малая длительность и нестационарный характер шумового сегмента, произносимого зачастую нечетко, затрудняют исследование и измерение его параметров. Этим можно объяснить наличие лишь единичных публикаций, описывающих признаки шумовых сегментов взрывных. Наиболее полная из них — работа Халле, Хьюза и Редли [162] — проводилась на речевом материале, содержащем односложные слова с взрывными в начальном и конечном положениях в сочетании с 5 гласными, а также с щелевыми. Дикторами были двое мужчин и одна женщина. Оpoznание взрывных по спектру шумового сегмента проводилось двумя ступенями. В первой ступени *t*, *d* и часть *k*, *g* различались от *p*, *b* и другой части *k*, *g* по соотношению энергии в полосах 0,7—10 кгц и 2,7—10 кгц с надежностью 95%. Во вто-

рой ступени t , d отличались от k , g по среднему уровню в полосах 0,3—10 кГц и 2—4 кГц с надежностью около 85%. Различение p , b и k , g проводилось по графику (где по одной оси отложены разности уровней двух наибольших спектральных пиков, $uf_1 - uf_2$, причем $f_2 > f_1$, а по другой оси — частота f_2) с надежностью ~82%.

В работе Пикеника [204] приведены осциллограммы взрывных, из которых следует, что длительность шумовых сегментов p и t составляет 18—25 мсек, k — 25—40 мсек. Длительность смычки — 20—250 мсек. Последнее значение имеет место при сочетании двух взрывных. Кривые распределения частот нулевых пересечений p , t , k в слогах типа СГ для одного диктора приведены Дошита [138]. Для шумового сегмента t максимум кривой расположен в диапазоне 5—9,6 кГц для сочетаний te , to и в диапазоне 3,3—5 кГц для ta . Частоты нулей для p (исключая сочетание pi) расположены довольно равномерно и охватывают почти весь диапазон спектра речи. k в сочетаниях с задними гласными характеризуется слабо выраженным максимумом около 1,2—1,5 кГц. Для k в слоге ka максимум на частоте 1,5 кГц выражен сильно. Максимумы кривых распределения частот нулей смещаются в области 3 и 4 кГц для k в сочетаниях ke и ki . О признаках взрывных, содержащихся в переходных сегментах гласных, и их роли при восприятии взрывных человеком уже указывалось в § 3 гл. 1 и § 4 этой главы. Отметим лишь, что для восприятия начальных взрывных более важен шумовой сегмент [159, 162].

Признаки шумовых сегментов начальных глухих взрывных

В ряде работ указывалось, что основная информация о взрывных заключена в огибающей сигнала, в частности на начальном участке [140, 146]. Тем не менее, как показали наши исследования восприятия клипированной речи (§ 2 гл. 2), в сигнале с постоянной амплитудой остается достаточно данных для восприятия взрывных. Это положение хорошо иллюстрируется габл. 8, 9, где приведены результаты испытаний на разборчивость глухих взрывных фонем при ограничении спектра сигнала частотой 7 кГц с некоторым ослаблением нижних частот.

Исследование признаков шумовых сегментов взрывных проводилось на речевом материале, состоящем из 38 односложных и двусложных слов, содержащих сочетания n , m , k со всеми гласными. Сигналы с выхода ФНЧ речевого тракта, описанного выше (рис. 1), поступали на шлейфный осциллограф типа МПО-2, где записывались на фотопленку. Запись производилась вибратором с собственной частотой 10 кГц. При 7 кГц отклонение луча вибратора составляло еще 30% от максимального значения. Пленка продвигалась со скоростью 1 м/сек. Каждая группа из 7—8 слов

произносилась большей частью 4 дикторами (двое мужчин и две женщины). Общее число дикторов составило 20 человек. Было записано также несколько слов с мягкими взрывными.

В результате анализа осциллограмм было установлено, что n , m , k , расположенные впереди гласных, можно различать по числу импульсов N на начальном участке шумового сегмента длительностью Δt и продолжительности шумового сегмента τ . Граница, отделяющая взрывной от последующего гласного, хорошо определялась по первому импульсу речевого сигнала длительностью $\theta \geq \theta_0 \cong 0,9$ мсек. Интересно отметить, что даже для сочетаний взрывных с a на границе, как правило, имелись импульсы длительностью $\geq \theta_0$, в то время как на протяжении a импульсы подобной длительности почти не встречались. В некоторых случаях, в соответствии с принятым определением τ , в продолжительность шумового сегмента включался интервал придыхания (аспирации).

На рис. 37, 38 приведены осциллограммы начальных участков слов *ток* и *тётя*. Параметры m в слове *ток* составляют $N = 29$ (при $\Delta t = 10$ мсек), $\tau = 12$ мсек. Граница между взрывным (шумовым сегментом) и гласным определилась по импульсу длительностью $\theta = 1,7$ мсек. Параметры m' ($N = 34$, $\tau = 65$ мсек) и характер звучания указывают, что эту фонему с большим основанием следует считать мягкой аффрикатой ζ' .

Результаты анализа начальных взрывных по признакам N (при $\Delta t = 10$ мсек) и τ приведены на рис. 39. Области, изображенные на рисунке, правильно классифицируют 95% m , 95% n и 75% k . Различение k от n в области $5 \leq N \leq 10$, $20 \leq \tau \leq 35$ мсек (где их параметры N и τ могут совпасть) достигается разделением по числу импульсов N' за время τ . В указанной области величина N' , как правило, для k больше, чем для n . Таким путем можно существенно улучшить процент распознавания фонемы k до значений, получаемых для n и m .

В зависимости от типа последующего гласного фонемы k разделяются по параметру N на два варианта. На рис. 39 начальные k , входящие в слоги *ка*, *кэ*, *кы*, показаны треугольниками, углом вверх, а k перед задними гласными *у*, *о* — такими же треугольниками, углом вниз. Пунктирная линия на рисунке, соответствующая значению $N = 9,5$, хорошо разделяет оба варианта k , даже лучше, чем исследованные признаки N и τ различают k от n и m . Необходимость выделения обоих вариантов твердого k зависит от назначения устройства. Так, если данные анализа звуков будут использованы для последующего синтеза речи, как в случае фонемного вокодера, нужно учитывать оба варианта k . В других случаях достаточно выделить одно твердое k . Параметры n и m мало зависят от типа последующего гласного.

Полученные признаки N и τ были проверены в устройстве для опознавания по ним начальных n , m , k со следующей блок-схемой

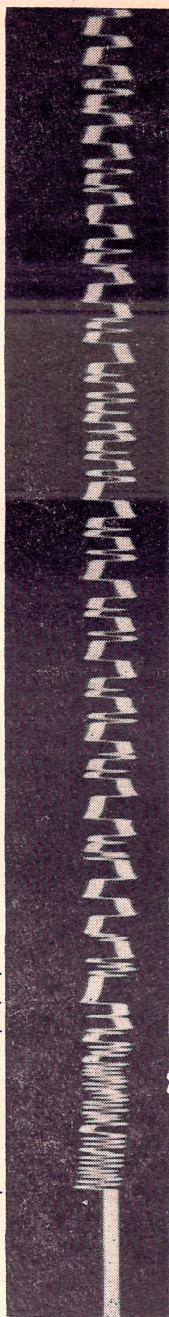


Рис. 37

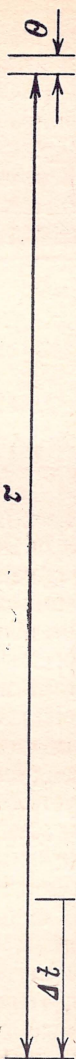
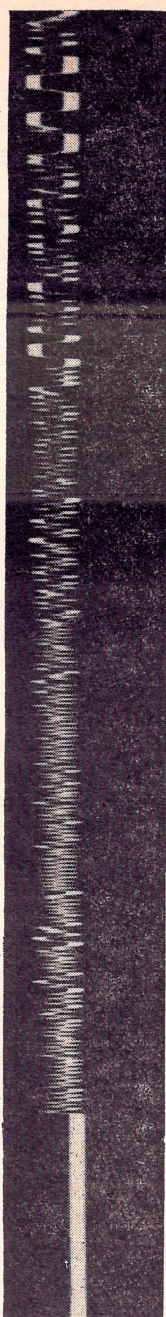
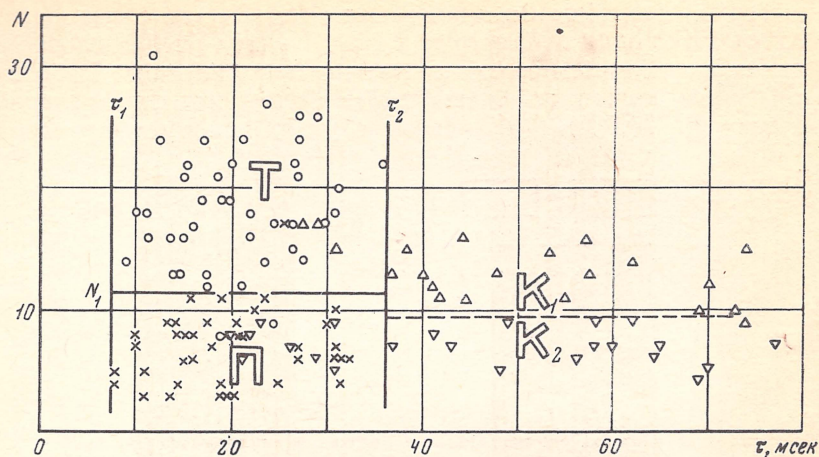


Рис. 38



Р и с. 39.

Результаты анализа начальных взрывных

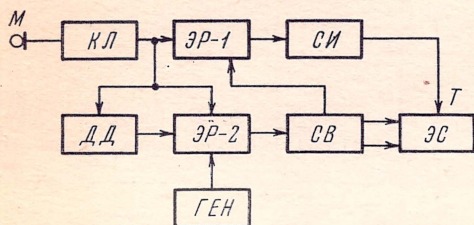
x — $п$; o — $т$; Δ — $к_1$ (перед $a, э, ы$); ∇ — $к_2$ (перед o, y)

(рис. 40). В исходном положении реле $\mathcal{EP}1$ находится в замкнутом состоянии, а $\mathcal{EP}2$ — в разомкнутом. Речевой сигнал после клипирования поступает одновременно через $\mathcal{EP}1$ на счетчик импульсов, на дискриминатор импульсов по длительности $ДД$ и на $\mathcal{EP}2$, включающее напряжение с генератора частотой f на счетчик времени. Таким образом, одновременно с подсчетом числа импульсов речевого сигнала производится отсчет пройденного времени. По истечении времени Δt счетчик времени через $\mathcal{EP}1$ запирает счетчик импульсов, и на последнем фиксируется число импульсов сигнала N на начальном участке. При поступлении на дискриминатор $ДД$ импульса, длительностью $\geq \theta_0$, реле $\mathcal{EP}2$ останавливает счетчик времени. На счетчике фиксируется второй признак — время τ . Признаки N и τ позволяют различить начальное $т$ не только от $п$ и $к$, но и от всех других фонем. Поэтому данное устройство было использовано для проверки влияния переспроса на повышение надежности опознавания начального $т$. При появлении $т$ (в сочетании с гласными) срабатывает элемент совпадения $\mathcal{ЭС}$.

Речевой материал состоял из 110 слов, произнесенных 12 лицами обоего пола. 23 слова начинались на $т$, остальные — на различные другие фонемы. Опознавание $т$ проводилось по граничным значениям признаков, несколько отличающихся от оптимальных: $\tau_1 = \Delta t = 7,2$ мсек, $\tau_2 = 29,4$ мсек, $\theta_0 = 0,7$ мсек, $N_1 = 8$. Время измерялось величинами, кратными 0,8 мсек.

На рис. 41, a показаны результаты испытаний. Каждая реализация $т$ характеризуется точкой на плоскости $N\tau$. Совпадения параметров двух, трех и четырех $т$ показаны кружочками с соот-

ветственно увеличенными диаметрами. При $\tau < \Delta t$ схема фиксировала только один параметр τ ; в этих случаях распределение по N дано произвольно. Всего было произнесено 1320 слов. В 276 словах, начинающихся на m , последнее было опознано правильно 214 раз (или 77,5%). В остальных случаях испытуемых просили

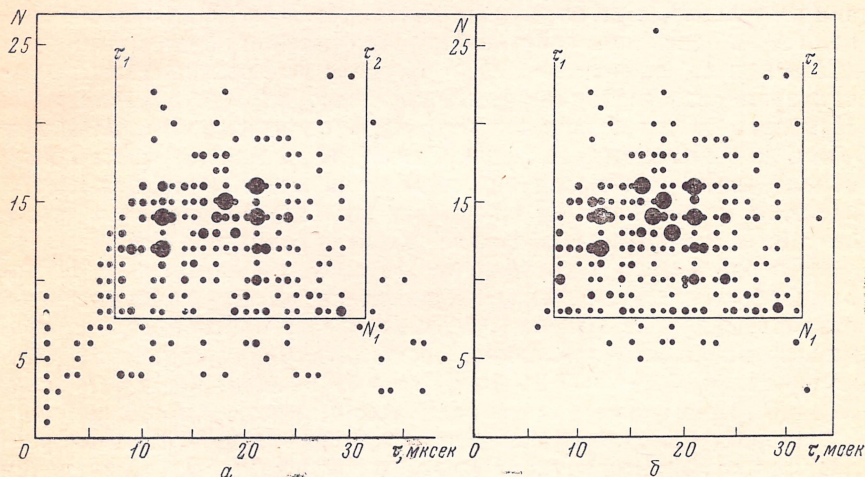


Р и с. 40

Блок-схема устройства для опознавания по признакам N и τ начальных n , m , k

M — микрофон; $КЛ$ — клиппер; $ЭР$ — электронное реле; $ГЕН$ — генератор; $ДД$ — дискриминатор импульсов по длительности; $СИ$ — счетчик импульсов; $СВ$ — счетчик времени; $ЭС$ — элемент совпадения

повторить слово более четко. В результате перепроса в 62 повторно сказанных словах m было правильно опознано 45 раз (или 72,6%). Надежность опознавания повысилась, таким образом, до 93,8%. Результаты испытаний с переспросом приведены на рис. 41, б. Несомненно, что применение второго переспроса для m , оставшихся неопознанными, привело бы к дальнейшему повышению надежности действия схемы. В остальных 1044 словах другие фонемы были показаны как m в 40 случаях (или 3,8%). Это были главным образом m' , n и k .



Р и с. 41

Решение задачи опознавания фонем по последовательности сегментов должно быть основано на использовании закономерностей формирования фонемных образов при восприятии речи человеком (см. § 3 гл. 1, а также работы Дукельского [35], Чистович и Кожевникова [110]). В общем случае необходимо выделять в непрерывном сигнале последовательности сегментов, каждая из которых характеризует одну определяемую фонему. Благодаря известной зависимости параметров данной фонемы от соседних эти последовательности оказываются существенно перекрывающимися во времени. Например, место образования согласного влияет не только на параметры переходного сегмента соседнего гласного, но и в небольшой мере и на его квазистационарную часть (см. табл. 12, а также [148]). В целях упрощения задачи можно включать в две последовательности (по которым опознаются две соседние фонемы) лишь переходные сегменты. Значительную трудность представляет выявление первого (начального) сегмента, содержащего существенную информацию об определяемой фонеме.

Решение рассматриваемой задачи осложняется наличием в последовательностях многих неопознанных и неполностью опознанных сегментов, что приводит также к отсутствию части границ между сегментами. В ходе дальнейших исследований речевых сигналов можно будет в какой-то мере восполнить эти пробелы. Тем не менее с подобного рода неопределенностями придется считаться и впредь. Как уже указывалось, человек разрешает эти неопределенности с помощью лингвистической информации. При автоматическом опознавании желательно увеличить роль акустических признаков за счет лингвистической информации, ввиду большого объема и громоздкости последней.

Вопросам опознавания фонем по последовательности сегментов посвящены пока немногие работы. В алгоритме Загоруйко [39] на первом этапе по совокупности признаков определяется принадлежность отрезков сигнала с одинаковой длительностью τ к одному из сегментов. Выдается столбец решений B_i , элементами которого являются вероятности принадлежности i -той реализации (отрезка) к q -тому сегменту. На основе опознавания обучающей последовательности сегментов составляется матрица, элементами столбцов A_s которой являются вероятности опознавания сегмента фонемы s как сегмента q . В эталон фонемы входит также характеристика ее длительности. Каждые τ мсек столбцы A_s проверяются на сходстве со столбцом B_i . В момент времени, когда функция сходства достигает максимального значения, вычисляются вероятности существования всех фонем принятого алфавита с учетом вероятностей диадных сочетаний фонем. После принятия решения в соответствии с максимальной вероятностью начинается заново формирование столбцов B_i , проверка столбцов A_s на сходство с B_i и т. д.

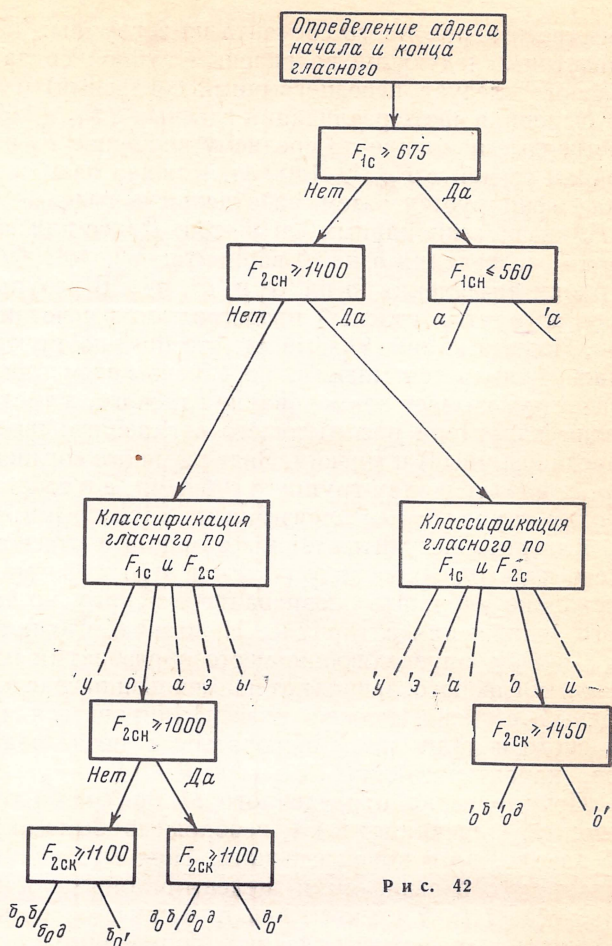
В основанных на этом алгоритме экспериментах, проведенных с использованием ЦВМ БЭСМ-6, $\tau = 14$ мсек [40]. В качестве признаков были приняты величины $\ln E$ и $\ln E/e_i$, где E — полная энергия, e_i — энергия на выходе одного из пяти октавных фильтров со средними частотами, начиная с 337,5 гц. Эксперименты подтвердили целесообразность использования информации о вероятностях появления сегмента и вероятностях сочетаний фонем. Сделан вывод о перспективности алгоритма. Однако количественных оценок эффективности примененного алгоритма не приведено.

Метод опознавания фонем путем непрерывно-группового преобразования динамических спектрограмм (типа «видимая речь») по характерным линиям (формантам) и характерным точкам (началам, максимумам и т. п.) описан в работах Файна и Сорокина. Принятие решения осуществляется по соответствию характерных точек предъявленной и эталонной спектрограммы после преобразования. При экспериментальной проверке с помощью ВМ БЭСМ-ЗМ вводилась информация о записанных в тихой комнате слогах *аба, ада, ага, ама, ана*, произнесенных 20 лицами по одному разу. Средняя надежность опознавания достигла 98% [80, 89, 90]. Метод опознавания фонем — по последовательности ненадежно опознанных отрезков сигнала равной длительности — изложен в работах Волошина и др. [14, 15]. Метод использует априорные вероятности сочетаний двух соседних сегментов.

При необходимости опознавания фонем по последовательности сегментов в реальном масштабе времени на машинах среднего класса (или класса типа БЭСМ-6, но работающей в режиме с разделением времени) не представляется возможным применять сложные и громоздкие вычисления. Кроме того, нужно учесть, что и в ближайшее время многие типы сегментов не смогут быть опознаны. В этих условиях следует, пожалуй, пока воздержаться от поисков «глобальных» методов опознавания фонем и рассмотреть отдельные важные случаи, которые должны быть и могут быть в ближайшее время реализованы.

Опознавание гласных

Определение признаков переходных и квазистационарных сегментов гласных дает возможность опознавать гласные по последовательности сегментов, что в принципе должно привести к повышению надежности опознавания не только неоднородных (дифтонгоидных) гласных, но и однородных. В п. 5 § 4 описывался метод опознавания «мягких» гласных, расположенных после глухих согласных и звонких взрывных, на речевом материале, использованном для исследования переходных сегментов гласных. Рассмотрим более полный алгоритм опознавания гласных в сочетаниях СГС, блок-схема которого приведена на рис. 42.



Р и с. 42

Исходные данные о произнесенном слове содержатся в записанной в оперативной памяти ЦВМ последовательности периодически вводимых признаков сигнала. До начала действия алгоритма определяются число гласных и ударный гласный. Первый блок определяет по признаку гласности и другим адрес начала и конца гласного. Эта задача к настоящему времени решена не для всех случаев, в особенности вызывает затруднение членение сочетаний гласных с сонорными. Второй блок выделяет часть реализаций *а* и *'а*, имеющих повышенное значение F_1 на протяжении квазистационарного сегмента. Эти реализации разделяются далее на твердые и мягкие по значению F_1 в конечной части начального переходного сегмента F_{1ch} . Гласные *'а* разделяются далее по F_2

конечного переходного сегмента на 2 группы: 'а^б (индекс «б» указывает на следующий за гласным губной согласный), 'а^д (после гласного следует переднеязычный согласный) и на 'а'.

Основная часть реализаций гласных с $F_{1c} < 675$ гц разделяется на твердые и мягкие по среднему значению F_2 в начальном переходном сегменте $F_{2сн} > 1400$ гц. Каждая из обеих групп гласных классифицируется затем на отдельные гласные по значениям F_1 и F_2 квазистационарных сегментов. Далее для примера показаны блоки разделения о на 4 варианта: $\text{о}^{\text{об}}$, $\text{о}^{\text{од}}$; $\text{о}^{\text{о'}}$; $\text{о}^{\text{дб}}$, $\text{о}^{\text{дд}}$ и $\text{о}^{\text{д'}}$, и блоки разделения 'о на 'о^б и 'о^д, 'о'. В результате работы данного алгоритма гласный представляется сочетанием трех сегментов. Первый из них указывает, что предшествующий гласному согласный является мягким, переднеязычным или губным. Второй сегмент указывает наименование гласного, а третий — пока 2 градации характера последующего согласного: твердый (губной или переднеязычный) и мягкий. Заднеязычные согласные в сочетаниях с у, о, входят в одну группу с губными, а в сочетаниях с а, э, ы — в одну группу с переднеязычными согласными.

Алгоритм не учитывает информации о гласном, содержащейся в его переходных сегментах. Учет этой информации позволил бы несколько улучшить опознавание гласного, но за счет значительного усложнения алгоритма. Напомним, что пороговые значения $F_{2сн}$ и $F_{2ск}$, определяющие место образования или наличие смягчения согласного, зависят от наименования гласного. Поэтому переходный сегмент должен классифицироваться при опознавании гласного и лишь затем включаться в последовательность сегментов, характеризующих согласный.

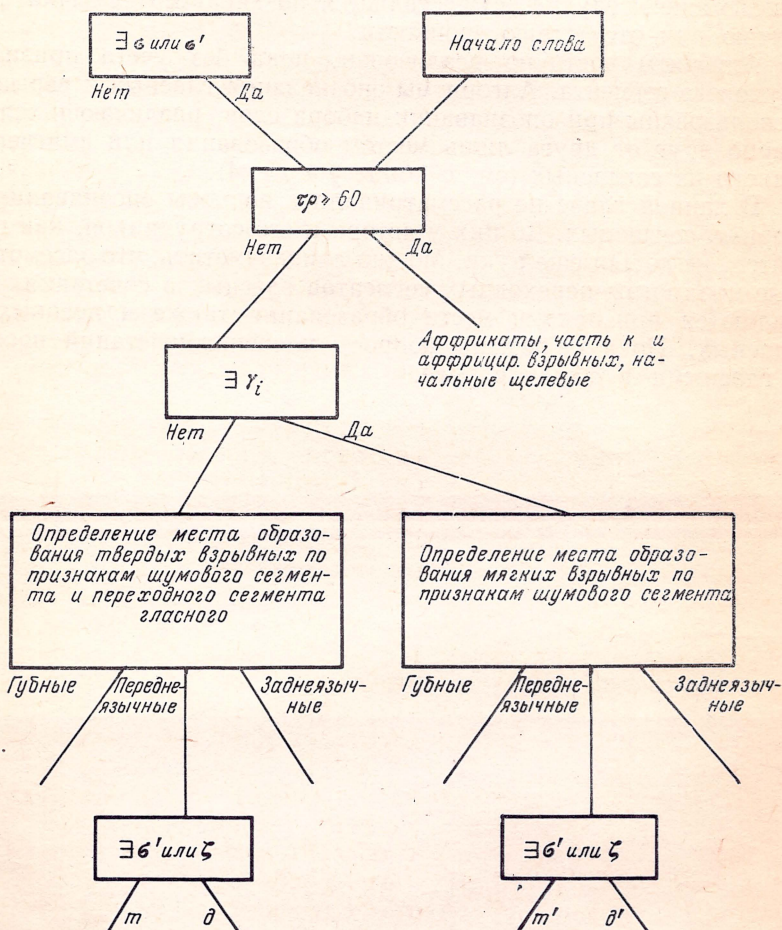
При проверке приведенного алгоритма, в отличие от эксперимента, описанного в § 4, необходимо было определять не только начало, но и конец гласного. Поэтому границы гласных в сочетаниях с взрывными и щелевыми определялись по признаку гласности (см. § 1 этой главы). Данные о групповых признаках и формантных частотах гласных, полученные в аналоговом устройстве, поступали каждые 20 мсек непосредственно на телетайпный вход ЦВМ «Минск-22». Слова произносились в помещении машинного зала. В речевом материале были широко представлены пары слов, различающихся друг от друга лишь местом образования или смягчением одного из согласных. Опознавание гласных по последовательности сегментов производилось ЦВМ в реальном масштабе времени.

Предварительные результаты проверки алгоритма показали, что начальный переходный сегмент гласных определяет место образования предшествующего твердого взрывного с надежностью около 90%. Для щелевых надежность снижается до ~80%. Напомним, однако, что для этого случая основной признак места образования фонемы содержится в параметрах квазистационарного сегмента щелевых. Что касается конечного переходного сег-

мента гласных, то разброс его параметров более значителен. С высокой надежностью можно определить место образования конечного *ч*. В общем случае необходимо также точнее находить конец гласного.

Опознавание взрывных

Алгоритм опознавания начальных взрывных в сочетаниях с гласными по смычке, шумовому сегменту и переходному сегменту последующего гласного приведен на рис. 43. Первый блок фиксирует наличие глухой или тональной смычки, указывающих на поступление взрывного согласного или аффрикаты. Второй блок



Р и с. 43

имеет два входа: второй из них для случая глухих взрывных в начале слова, для которых смычка совпадает с предшествующей слову паузой и не может поэтому являться признаком. По длительности шумного сегмента тр группа взрывных отделяется от аффрикат, части аффрицированных взрывных и других звуков. В следующем блоке взрывные разделяются на твердые и мягкие по характеру переходного сегмента соседнего гласного с использованием блоков, описанных в предыдущем алгоритме. Далее, твердые взрывные разделяются по месту образования по признакам шумового сегмента и переходного сегмента соседнего гласного, а мягкие взрывные только по признакам шумового сегмента. Эти два блока представлены схематически. Разделение взрывных на звонкие и глухие осуществляется по характеру смычки и наличию или отсутствию звонкости.

Алгоритм частично реализован, пока без учета признаков шумового сегмента. Алгоритмы опознавания гласных и взрывных использованы при опознавании набора слов, различающихся парно друг от друга лишь местом образования или смягчением одного из согласных (см. в конце § 4 гл. 4).

В данной главе не рассматривались вопросы опознавания сонорных согласных. До них у автора и его сотрудников, как говорится, не дошли еще руки. Можно лишь заметить, что рассмотренные параметры переходных сегментов гласных в сочетаниях СГС являются признаками места образования также и носовых согласных. Пока не решены вопросы членения сочетаний носовых с гласными в общем случае.

ОПОЗНАВАНИЕ ОГРАНИЧЕННОГО НАБОРА СЛОВ

§ 1. О признаках речевого сигнала на уровне слов

В настоящее время, а также и в ближайшем будущем опознавание слов придется осуществлять лишь по части входящих в слово фонем и сегментов, о некоторых из которых к тому же имеется неполная информация (например, только о способе образования). Кроме того, из-за различий в произношении, рассмотренных в § 2 гл. 1, а также вследствие неточностей и погрешностей при выделении признаков последовательности сегментов, отображающие одно и то же слово, получают с большим разбросом. Процедура опознавания слов по последовательности фонем и сегментов в целом становится громоздкой, в связи с чем возникла необходимость применения различного рода словесных признаков. В общем случае словесные признаки должны отражать наличие, местоположение и взаимное расположение фонем и сегментов в слове, положение ударного гласного, а также соотношения по длительности определенных сегментов и слова в целом.

Число и тип применяемых словесных признаков зависят от размеров словаря и типа анализатора. При опознавании 10—20 слов основную роль играют групповые признаки, позволяющие различать звуки по способу образования. Для разделения большинства слов набора этого оказывается достаточным. Для разделения остальных слов, совпадающих по последовательности используемых групповых признаков, приходится применять признаки места образования или соотношения длительностей сегментов. Начиная с нескольких десятков слов становится необходимым определять положение и принадлежность некоторых опорных фонем, выделяемых с большей надежностью, в первую очередь ударных гласных и длительных шумных согласных. Словесные признаки будут рассмотрены подробнее в ходе дальнейшего изложения. Ограничимся здесь их примерным перечнем с указанием обозначений.

а) Признаки наличия и количества определенных сегментов и фонем в слове.

Эр — наличие шумного сегмента (фонемы).

Эс — наличие фонемы с.

пγ — число гласных.

$n\sigma$ — число глухих смычек.

ni — число фонем $и$.

Аналогичным образом обозначаются признаки наличия и количества других сегментов и фонем.

б) Признаки местоположения сегментов и фонем.

$\dot{\gamma}_1, \gamma_2, \gamma_3, \gamma_y, \gamma_z$ — соответственно первый (являющийся также ударным), второй, третий, предпоследний и последний гласные.

$\sigma_1, \sigma_2, \sigma_z$ — первая, вторая и последняя глухие смычки.

ρ_n, ρ_c, ρ_k — начальный, средний и конечный шумные.

Например, ρ_n имеется в наличии, если слово начинается шумовым сегментом. В этом случае ρ_n совпадает с ρ_1 .

в) Признаки взаимного расположения.

Порядок следования двух или более сегментов указывается порядком следования их обозначений, например:

$\sigma\rho_k$ — сочетание смычки и шумного сегмента ρ_k .

$\dot{\gamma}_1\rho_1$ — сочетание ударного гласного $\dot{\gamma}_1$ и ρ_1 .

$\sigma - {}^b y$ — сочетание смычки и гласного ${}^b y$, разделенных другими сегментами.

г) Признаки определенных сегментов, фонем и звуко сочетаний в слове.

$F_1\dot{\gamma}_1$ — первая форманта квазистационарного сегмента гласного $\dot{\gamma}_1$.

$N\rho_k$ — число интервалов между нулями в квазистационарной части ρ_k .

$\cup F_1\zeta_1$ — спад F_1 с последующим подъемом в звуко сочетании ζ_1 .

д) Признаки длительности.

T — длительность слова.

$t\gamma_1$ — длительность гласного.

$t\sigma\gamma_1$ — длительность смычки, расположенной до γ_1 .

$t\sigma\gamma_y\gamma_z$ — длительность смычки, расположенной между гласными γ_y и γ_z .

$\rho_c t\rho_k$ — длительность промежутка от конца ρ_c до начала ρ_k .

$Ht\gamma_1$ — длительность промежутка от начала слова до γ_1 .

Остановимся несколько на опорных фонемах, к которым можно привязывать остальные элементы слова и по которым целесообразно разделять набор опознаваемых слов на несколько групп. В качестве опорной фонемы естественно принять ударный гласный, что и было сделано в исследовании Голда [157] (описанном в следующем параграфе) и при опознавании некоторых групп слов в наших работах. Однако при выделении ударного гласного, характеризуемого большей интенсивностью и длительностью, встречается ряд трудностей. Например, интенсивность неударного a зачастую превышает интенсивность ударного $и$. Определение длительности гласных в сочетаниях с сонорными затрудняется из-за отсутствия достаточно надежного метода их расчленения, в связи с чем при-

ходится пока рассматривать часть этих сочетаний как «машинные» гласные. Подобный неударный «машинный» гласный может по длительности превышать ударный. В качестве опорных фонем оказалось удобным использовать длительные шумные *ш, с, з, ц, ч* (в словах, где они имеются).

§ 2. Основные направления исследований

В связи с тем, что ближайшей разрешимой задачей по проблеме автоматического опознавания речи является разработка машин, опознающих ограниченный набор слов, решению этой задачи посвящены десятки исследований. В большинстве из них объектом опознавания являются 10 цифр (иногда еще несколько других слов), произносимых на русском языке [42, 47, 52, 82, 102], английском [132, 136, 137, 143, 152, 181, 203, 214], японском [194, 222, 230], немецком [117, 182, 193, 226], итальянском [156].

Рассмотрим вначале машины аналогового типа. Действие машины, описанной Дэвисом, Бидульфом и Балашеком [132], основано на сравнении формантных кривых последовательности звуков слова (цифры) на плоскости $F_1 F_2$ с 10 эталонными кривыми. Частоты F_1 и F_2 определялись подсчетом числа нулей после клипирования сигнала в диапазонах ниже и выше 900 *гц*. Надежность опознавания 10 цифр составляла 98% при настройке схемы на голос говорящего и уменьшалась до 50—60%, когда цепи машины не подстраивались под другого диктора.

В машине Дадли и Балашека [143] речевой сигнал в диапазоне до 2950 *гц* делился на 10 полос, первая из которых была шириной 250 *гц*, а остальные — по 300 *гц*. Речевой сигнал, соответствующий произнесенной цифре, по соотношению напряжений в 10 частотных полосах преобразуется в последовательность фонетических элементов длительностью 41,7 *мсек*, грубо соответствующих 6 гласным и 4 согласным. По количеству и наименованию фонетических элементов определяется произнесенная цифра. Порядок, в котором появляются элементы, во внимание не принимается. Машина работает «почти отлично» при настройке на определенный голос. Другие лица того же пола после некоторой тренировки могут добиться правильного действия машины примерно в 90% случаев.

Словарь опознающей машины, разработанной в Тбилиси (Какауридзе, Доценко [48], Какауридзе [47]), содержит кроме 10 цифр еще 10 слов: *плюс, минус, пробел, вперед, назад, направо, налево, быстро, медленно, стоп*. С помощью речевых команд осуществляется печатание цифр и управление движением модели тележки. Речевой сигнал в диапазоне 200 *гц* — 11 *кгц* анализировался 7 полосовыми фильтрами с дифференцированием и клипированием сигнала в каждой из полос до фильтрации. Признаками сигнала являются фиксируемые на выходе частотных каналов каждые 50 *мсек* три значения уровня: нулевой (малый), сред-

ний и высокий. В качестве словесных признаков при построении дешифрирующих схем (отдельно для каждого слова) были использованы: наличие и число определенных признаков в слове, порядок их следования во времени, скорость изменения интенсивности сигнала, соотношения энергии сигнала в некоторых каналах. Надежность опознавания достигала 98% для лиц, на которых был построен один из трех имевшихся дешифраторов.

Дальнейшие усилия исследователей были направлены на получение словесных признаков, зависящих в меньшей степени от индивидуальных особенностей говорящих. Машина Нагата, Като и Чибя [194, 230] разрабатывалась с учетом возможности ее применения в дальнейшем в телефонной связи для набора номера абонента. Упор был сделан поэтому на звонкую часть сигнала. После прохождения сигнала через полосовые фильтры определялись F_1 , F_2 , интенсивность и основной тон. Опознавание производилось по 8 словесным признакам: числу звонких сегментов в слове $n\zeta$, значению F_2 во втором звонком сегменте ζ_2 , наличию (или отсутствию) глухого согласного в начале слова $\mathcal{E}_{p_n}$ и пяти параметрам первого звонкого сегмента ζ_1 : положению F_1 и F_2 в начале, 100 мсек после начала и в конце сегмента, крутизне F_2 в начале и характеру изменения F_1 в течение сегмента. Число градаций признаков от двух до 13. Решение принимается на вероятностной основе по правилу Байеса с помощью матрицы, где сопротивления, соединяющие шины признаков и слов, соответствуют обратным вероятностям того, что при данном значении признака i будет цифра j . Надежность действия машины составила 99,7% для 1000 слов 1 лица и 97,9% для 1000 слов 20 лиц (мужчин)..

Решение на вероятностной основе с помощью матрицы сопротивлений принимается также в машине, описанной Книппером и Петровым [52]. Речевой сигнал проходит через 5 октавных фильтров и параллельно через цепи, измеряющие усредненную плотность нулей ρ_{τ} (в том числе после дифференцирования и интегрирования сигнала). В качестве двоичных словесных признаков были приняты отношения различным образом полученных плотностей нулей и отношения энергии в полосах в начале и внутри слова, а также число одинаковых признаков в слове. Надежность опознавания при проверке на 25 дикторах составила 96%.

В «Шубоксе» Дерша [137], предназначенном для работы с суммирующей машинкой, 16 слов (включая цифры) по типу расположения звонких и шумных сегментов классифицируются на 4 группы. Дальнейшее разделение производится по уровню щелевых и другим неуказанным признакам. Началом отсчета слова служит появление асимметричного сигнала (разности огибающих отрицательных и положительных полувольт) первого звонкого сегмента. Если слово начинается с глухого звука, информация о нем хранится до поступления звонкого сегмента. Название «Шубокс» подчеркивает малые габариты устройства (Shoebox —

коробка для обуви). По более поздним данным [130] 15 слов в «Шубоксе» разделяются по сочетаниям звонких, щелевых и взрывных на 12 групп (с учетом уровня щелевых), причем лишь в двух группах необходимо привлечь дополнительные признаки. При проверке с участием 7 мужчин и 7 женщин надежность опознавания составила 96,8% (938 ошибок на 29400 реализаций).

При опознавании 10 наименований немецких цифр (Куш [182]) использованы два бинарных признака e_n и e_v — наличие энергии в диапазонах до 1 кГц и выше 2 кГц. Слова кодируются пятизначным кодом, средний элемент которого соответствует звукам a , e , i , l , r , имеющимся в каждой цифре и содержащим оба признака e_n и e_v . Остальные элементы кода представляют звуки, содержащие e_n (задние гласные и носовые) или e_v (щелевые). В случаях, когда средний элемент кода, содержащий e_n и e_v , соответствует начальному (конечному) звуку слова, первые (последние) два элемента кода равны нулю. Надежность опознавания цифр, произнесенных 37 мужчинами и 37 женщинами, составила 87%, с нормализацией по уровню — 93%. Размеры устройства $10 \times 17 \times 35$ см³. Фонетическая суммирующая машинка Мусмана и Штейнера [193] опознает 14 слов. Решение о принадлежности слова принимается по 30 двоичным признакам (3 отсчета на слово по 10 частотным каналам, охватывающим диапазон 200—6600 Гц). Надежность действия на одно лицо — 97%, для 10 других лиц — 87%. На разброс признаков влияют громкость и темп произношения.

Рассмотрим далее исследования с использованием ЦВМ для принятия решения при опознавании и в некоторых случаях для анализа. В работе Форджи [152] каждый отрезок сигнала (прошедшего через 35 полосовых фильтров) длительностью 16,7 мсек классифицировался ЦВМ ТХ-2 как принадлежащий к гласному (с номером, зависящим от F_2), глухому или звонкому согласному, взрывному. Однородные отрезки объединялись в сегменты с двумя градациями по длительности. Решение о принадлежности слова принималось по последовательностям сочетаний трех сегментов, формировавшихся со сдвигом на один сегмент. За произнесенную цифру принималось слово, которое оказывалось возможным для всех сочетаний троек сегментов данного слова. Надежность опознавания цифр, произнесенных 10 лицами (7 мужчин, 3 женщины), составила 98%.

В исследовании Шольца и Бекиса [214] речевой сигнал вводился в ЦВМ через преобразователь аналог — код. С помощью смоделированной системы из 40 полосовых фильтров (в диапазоне до 8 кГц) каждые 10 мсек определялось распределение энергии по спектру. По отношению энергии в определенных полосах отрезки сигнала классифицировались на 11 категорий условных гласных методом многомерных статистических решений и на 3 категории согласных (взрывные, сильные и слабые щелевые). Принятие решения производилось путем сравнения полученной последователь-

ности сегментов с эталонными. Надежность опознавания цифр на материале 50 дикторов обоего пола составила 97%.

Метод опознавания слов (цифр) по распределению длительностей интервалов между нулями клипированного сигнала описан Зайцевым и Тимофеевым [42]. Каждое слово выражается матрицей с 10 столбцами (по числу градаций длительностей интервалов) и количеством строк, равным числу отсчетов, кратным 20 мсек, для наиболее продолжительного слова. Для коротких слов недостающее количество строк заполняется нулями. Решение принималось ЦВМ по максимуму взаимной корреляции произнесенного слова с эталонами, хранящимися в памяти машины. Надежность опознавания 10 цифр четырьмя дикторами обоего пола превысила 98%.

Исследования по опознаванию слов, состоящих главным образом из ограниченного числа фонем, проводили Дадли [142], Фрай и Динес [135, 155], Гумецкий, Дергач и др. [31]. В первой из этих работ исследование проходило в рамках моделирования ограниченного фонемного вокодера. Опорные фонемы *i*, *I*, *e*, *a*, *o*, *u*, *s*, *f*, *r*, *n* опознавались в соответствии с процедурой, изложенной при рассмотрении [143]. Сигналы, соответствующие опознанным фонемам, поступали в синтезатор и далее в громкоговоритель. Цифры и детские стишки были поняты слушателями. Разборчивость цифр произнесенных лицом, на которого была настроена опознающая часть, составила 98% (при двух слушателях). Однако при прослушивании цифр и 37 других употребительных односложных слов разборчивость снизилась до 55,5%.

В машине Фрая и Динеса [135, 155] спектральный анализ производился с помощью 18 фильтров, охватывающих диапазон 160 гц — 8 кГц. Опознавались слова, состоящие из *i*, *u*, *a*, *e*, *k*, *t*, *s*, *ʃ*, *f*, *z*, *m*, *n*, *l*. Признаками фонем являются максимумы произведений напряжений двух частотных полос. Согласные, имеющие похожие спектры, различаются по длительности или по уровню. При опознавании слов использовалась информация о вероятности появления отдельных двухфонемных сочетаний. При настройке машины на голос говорящего машина узнавала 60% фонем, а при добавлении указанной лингвистической информации — 72%. Последнее значение снижалось до 45% при произнесении тех же слов двумя другими лицами.

В работе Гумецкого, Дергача и др. [31] опознавались слова, состоящие из гласных *y*, *o*, *a*, *a'*, *э*, *э'*, *ы*, *и*, твердых и мягких щелевых *с*, *ф*, *х*, *ш*. После спектрального анализа и предварительной сегментации из сигнала выделялись огибающая спектра, признак звонкости и признаки, соответствующие спектральным полосам с максимальными различиями для противопоставляемых пар фонем. Фонемная принадлежность определялась в двух классификаторах фонем по звонкости и спектральной огибающей, по

звонкости и полосным признакам. Для словаря из 73 слов, содержащих только сочетания щелевых и гласных, надежность опознавания составила 93% (при двух дикторах — мужчине и женщине). Опознавание производилось с помощью ЦВМ «Минск-22».

Опознаванию нескольких десятков слов посвящены пока немногие работы. В статье Кинга и Тьюниса [180] описан эксперимент по опознаванию 30 слов с помощью ЦВМ IBM 1620-П. Речевой сигнал проходит через 15 полосовых фильтров, охватывающих диапазон 140—4500 *гц*. Информация о пиках (локальных максимумах) огибающей спектра вводится в ЦВМ. Входные данные преобразуются в различные пространства измерений, удобные для применения к ним линейных решающих функций. Надежность опознавания для одного диктора при обучении на его голос составила 99,5% при 15 словах и 98,7% при 30 и резко снижалась для других лиц.

В заметке [188] сообщаются краткие сведения о системе *SIDS*, предназначенной для связи человека с ЦВМ и опознающей не только набор в 30 слов, но и говорящего из числа примерно 30 определенных лиц. Анализ речевого сигнала выполняется с помощью гребенки полосовых фильтров при частоте отсчета 200 *гц*. Опознавание производит ЦВМ.

Наиболее обширный словарь, из числа известных нам зарубежных исследований, содержится в работе Голда [157]. В числе опознаваемых 54 слов — 10 цифр, математические термины, служебные слова для управления ЦВМ. Речевой сигнал поступает на 16-канальный анализатор спектра, работающий в диапазоне 180—3000 *гц*, и параллельного на блоки выделения основного тона и звонкости. Отсчеты спектральных каналов вводятся каждые 2,5 *мсек* (основного тона и звонкости через 5 *мсек*) в ЦВМ ТХ-2. Слово разделяется на звонкие и глухие отрезки. Членение звонких отрезков на сегменты производится методом, описанным в § 2 гл. 3. Анализу подвергалась лишь часть слова, включающая ударный сегмент (определяемый по наибольшей энергии) и не более чем по 2 соседних ему сегмента с каждой стороны. В качестве 15 словесных признаков использовались относительные длительности сегментов τ_i/T , наличие пауз $\Delta\sigma$, отношения энергии в нескольких определенных каналах к общей энергии e_i/E в ударном и предупредительном сегментах и средние значения F_1 , F_2 , dF_2/dt , $F_{2н}$ — $F_{2к}$ в ударном сегменте. Решение о принадлежности принимается путем сравнения определяемого слова со всеми словами, по тому, как значения его признаков укладываются в диапазон значений каждого признака для каждого слова (попадает в диапазон, выше или ниже). Надежность опознавания 54 слов, произнесенных каждое 10 мужчинами, составила 86,3%. Слова предварительно записывались в сравнительно тихой комнате на магнитофон.

В работе Осадчего [70] речевой сигнал анализируется аналоговым устройством с помощью 9 полосовых фильтров в диапазоне

150 гц — 12 кгц. Каждый отрезок длительностью 12,5 мсек классифицируется как один из 41-го фонетического элемента, включая глухую и тональную смычки. Слово делится на 10 участков. Эталон для слова представляет собой перечень элементов, разрешенных для каждого данного участка. Элементы каждого участка сравниваются с эталоном. Решение принимается по эталону, в котором разрешенным оказывается наибольшее суммарное число элементов. Для 32 слов, произнесенных 9 лицами (7 мужчин, 2 женщины), надежность опознавания составила 95%. Время опознавания на ЦВМ «Урал-11б» составляет 3 сек.

Эксперимент по опознаванию 168 слов одним диктором с помощью ЦВМ БЭСМ-6 описан Величко и Загоруйко [11]. Словарь включал термины языка «Алгол-60» и α -транслятора, названия математических функций. Речевой сигнал пропусклся через систему из пяти полосовых фильтров со средними частотами: 112,5, 450, 900, 1800 и 7200 гц. Каждые 14 мсек подсчитывались энергия в каждой полосе e_i и общая — E . Признаками сегментов служили значения $\ln(E/e_i)$. Слово описывалось последовательностью сегментов. Для принятия решения слово сравнивалось приближенным методом со всеми 168 словами-эталонами. 16 наиболее похожих на них отбирались для точного определения меры сходства с предъявленным словом. Надежность опознавания 168 слов, произнесенных диктором-мужчиной в тихой комнате, составила 95,3%. Время опознавания около 10 сек. Позднее словарь был расширен до 203 слов при сохранении той же надежности.

§ 3. Опознавание нескольких слов по последовательностям сегментов

В исследованиях последних лет число опознаваемых слов медленно, но неуклонно растет, достигая к концу 1969 г. нескольких десятков для многих дикторов и порядка двухсот слов при настройке системы на голос говорящего. Тем не менее исследования и разработки устройств, опознающих несколько слов, продолжают. Подобные устройства, будучи компактными и достаточно надежными, могут найти практическое применение в некоторых случаях.

В излагаемых ниже экспериментальных исследованиях речевой сигнал, соответствующий произнесенному слову, рассматривается состоящим из последовательности сегментов. Один сегмент может соответствовать двум и более звукам (при отсутствии признаков для их разделения или на стыках между звуками). С другой стороны, один звук (взрывной, аффриката) может быть представлен двумя и более сегментами.

Рассмотрим зависимости между признаками, сегментами и словами для некоторого набора слов.

Пусть имеется ряд сегментных признаков $a_1, a_2, \dots, a_k, \dots, a_n$. Можно написать, что $a \in A$, где A — множество признаков.

Отрезок сигнала, обладающий совокупностью признаков $a_1, a_2, \dots, a_m$, отображает сегмент b_k . $b \in B$, где B — множество сегментов.

В отдельных случаях сегмент может характеризоваться только одним признаком или даже отсутствием признаков (при глухой смычке).

Однократному произнесению какого-либо слова будет соответствовать последовательность сегментов во времени $c_{ak} = b_1, b_2, \dots, b_m$. Можно принять, что $c_a \in C_a$, $C_a \subset C$, где C_a — подмножество последовательностей, соответствующих реализациям одного слова, а C — множество последовательностей, соответствующих реализациям всех слов набора.

Для заданного набора слов имеем ряд подмножеств $C_a, C_b, \dots, C_k \dots C_n$, каждое из которых представляет возможные последовательности сегментов одного из слов набора. Число элементов этих множеств (вариантов реализаций слов) следует уменьшить путем объединения эквивалентных сегментов (отличающихся нехарактерными для них признаками) и отбрасывания коротких звонких сегментов, что соответствует операции препарирования объектов при опознавании зрительных образов [93].

В результате анализа полученных последовательностей для заданных слов могут встретиться три случая.

а) Все последовательности сегментов, входящие в C , различаются друг от друга, т. е. каждое слово набора имеет варианты, присущие только этому слову. Опознавание каждого слова набора может быть принципиально осуществлено без каких-либо затруднений.

б) Имеются случаи совпадения последовательностей сегментов разных слов набора. Однако общий процент неопределенностей, вносимых при опознавании, мал. В подобных случаях следует принимать решение на вероятностной основе (с допущением некоторого числа ошибок) или осуществлять переспрос. Как показывают экспериментальные данные (см. § 6 гл. 3, а также [101]), большая часть повторных произнесений приводит к правильному опознаванию.

в) Число совпадающих последовательностей велико, причем в их число входят часто встречаемые варианты. Применение переспроса в этом случае не даст необходимого эффекта. Для устранения неопределенностей необходимо применить дополнительные признаки или заменить часть слов заданного набора другими.

Опознавание нескольких слов по усредненной плотности нулей ρ_τ

С целью осуществления ввода голосом двоичной информации была исследована возможность весьма надежного опознавания трех слов по функции $\rho_\tau(t)$, произносимых разными лицами полным стилем — голосом и шепотом. Функция $\rho_\tau(t)$ является однопо-

лярной, и на некоторых отрезках времени (перед смычными фонемами) ее значение близко или равно нулю. В качестве признаков, различающих небольшое число слов, можно принять следующие параметры: число импульсов n , составляющих кривую $\rho_\tau(t)$, и максимальные значения $\rho_{\tau\max}$ в каждом из n импульсов.

Для поставленной задачи были заданы слова *один*, *ноль*, *стоп*. На рис. 44 приведены осциллограммы признака полной энергии E_τ (вверху) и ρ_τ (внизу) для указанных выше слов. Постоянная времени усреднения $\tau = 10$ мсек. Значения признаков отложены по вертикали вниз. Для слова *ноль* кривая ρ_τ при нормальном произношении состоит из одного импульса $n = 1$. Для слова *один* $n = 2$. Для слова *стоп* в общем случае $n = 3$. У большинства лиц взрыв (шумовой сегмент), соответствующий фонеме n , выражен нечетко и благодаря ее малой длительности, а также усреднению ρ_τ достигает на этом отрезке времени весьма малого значения. Поэтому в дальнейшем третий импульс ρ_τ слова *стоп* нами не учитывался.

На рис. 45 и 46 приведены распределения величин $\rho_{\tau\max 1}$ и $\rho_{\tau\max 2}$ для заданных слов, произнесенных голосом и шепотом 25 лицами обоего пола. Речевой тракт состоял из микрофона ДЭМШ, усилителя, ограничителя по минимуму и максимуму и кип-реле, нагруженного на интегрирующий контур с $\tau = 40$ мсек. Увеличение τ снизило существенно неравномерность кривой ρ_τ по сравнению с приведенной на рис. 44. Масштаб кривой ρ_τ определялся по синусоидальному сигналу с частотой f , для которого $\rho(t) = 2f$. Схема обеспечила линейность ρ_τ для частот до 7 кГц.

Для слова *ноль* (рис. 45) представлено одномерное распределение $\rho_{\tau\max 1}$. На оси ординат отложено число слов, имеющих данную величину $\rho_{\tau\max 1}$. В целом кривая охватывает диапазон 1200—4400. Слова, произнесенные шепотом, имеют в среднем более высокие значения $\rho_{\tau\max}$, причем значения выше 3000 относятся большей частью к женскому шепоту.

На рис. 46 приведено распределение обоих максимумов ρ_τ для слов *один* и *стоп*. Для слова *один*:

$$1000 \leq \rho_{\tau\max 1} \leq 4200, \quad 1400 \leq \rho_{\tau\max 2} \leq 9600.$$

Нижние значения обоих максимумов относятся к мужскому произнесению голосом. Верхние значения относятся большей частью к шепотной речи. Слово *один* произносилось большинством дикторов как [адзин]. Большой разброс значений $\rho_{\tau\max 2}$ объясняется различиями в степени аффрицирования d' (точнее, $d\bar{z}$), особенно заметными при шепоте.

Диапазон значений параметров для слова *стоп* составил

$$4600 \leq \rho_{\tau\max 1} \leq 10\,000, \\ 800 \leq \rho_{\tau\max 2} \leq 4200.$$

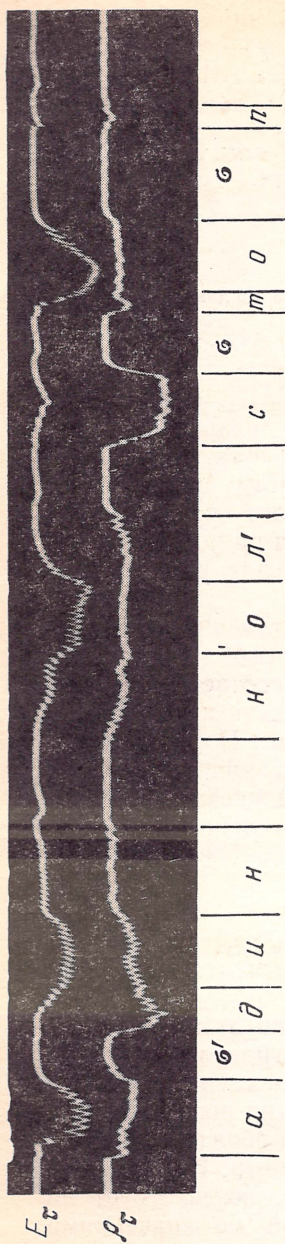


Рис. 44

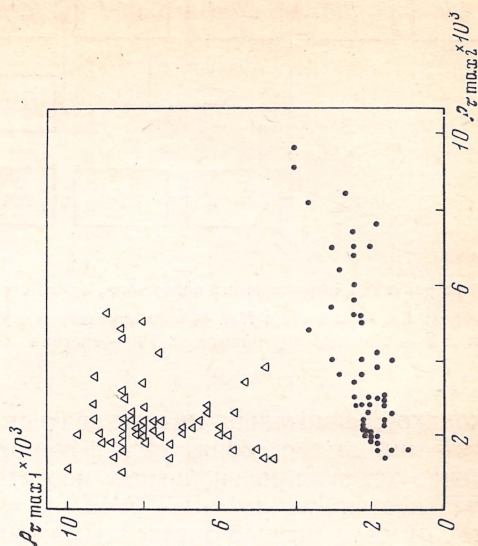


Рис. 46
Распределение величин $\rho_{\tau \max 1}$ и $\rho_{\tau \max 2}$
 Δ —стой, $\circ$ —один

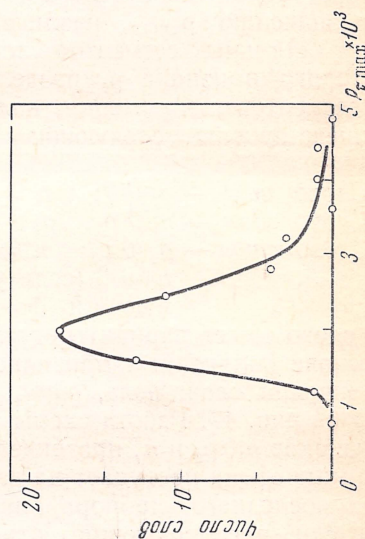
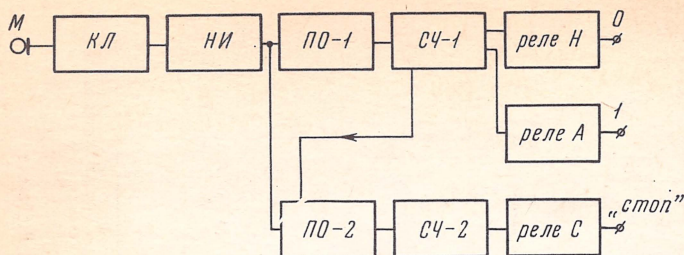


Рис. 45



Р и с. 47

Блок-схема устройства для различения слов: ноль, один, стоп

М — микрофон; КЛ — клиппер; НИ — нормализатор импульсов по длительности с интегратором; ПО — пиковый ограничитель; СЧ — счетчик

Несколько сравнительно низких значений $\rho_{\text{тmax1}}$, если учесть, что первый импульс $\rho_{\text{т}}$ образуется фонемой *с*, можно объяснить недостаточно четким произношением последней, а также меньшей чувствительностью микрофона ДЭМШ к частотам выше 5 кГц. Отдельные отрезки сигнала оказались ниже порога срабатывания, что привело к уменьшению $\rho_{\text{т}}$. Все же, несмотря на подобные случаи, области распределения максимумов $\rho_{\text{т}}(t)$ для рассматриваемых трех слов не перекрываются.

Исходя из полученных данных, можно установить два сегментных бинарных признака для различения слов: *ноль*, *один*, *стоп*. Порог срабатывания первого из них ρ_1 соответствует граничному значению $\rho_{\text{тmax}}$, различающему глухую и тональную смычки от остальных сегментов слов *ноль* и *один*. Пороговое значение второго признака ρ_2 разделяет область значений $\rho_{\text{тmax1}}$ первых сегментов *стоп* и *один*, *ноль*. При этом указанные слова будут представлены следующими последовательностями сегментов.

Слово *ноль* — $\rho_1 0 0$; $n = 1$.

Слово *один* — $\rho_1 0 \rho_1$, $\rho_1 0 \rho_2$; $n = 2$.

Слово *стоп* — $\rho_2 0 \rho_1$; $n = 2$; $\rho_2 0 \rho_1 0 \rho_1$,

$n = 3$ (если учитывать сегмент, соответствующий фонеме n при $\rho_{\text{тmax3}} \geq \rho_1$).

Каждое слово имеет варианты, присущие только этому слову.

На основе полученных признаков было построено устройство для различения слов: *ноль*, *один*, *стоп*. Блок-схема устройства показана на рис. 47. Работа схемы, основанной на использовании трех признаков ρ_1 , ρ_2 и n , протекает следующим образом. Речевой сигнал с микрофона поступает на клиппер и рометр. Напряжение с выхода последнего, пропорциональное $\rho_{\text{т}}(t)$, воздействует на ограничители, преобразующие отрезки кривой со значениями $\rho_{\text{т}} > \rho_1$ и $\rho_{\text{т}} > \rho_2$ в прямоугольные импульсы. Счетчик 1 фиксирует один или два импульса с уровнем $a\rho_1$. При $n = 1$ с задержкой

в 360 мсек срабатывает реле 0, при $n = 2$ — реле 1 (с задержкой 50 мсек). Счетчик 2 фиксирует импульс напряжения br_2 при $r_t > r_2$, по окончании которого через 25 мсек срабатывает реле стоп. После поступления первого импульса на счетчик 1 последний блокирует вход ограничителя 2, в результате чего при слове *один*, где иногда $r_{tmax2} > r_2$, ложного срабатывания реле *стоп* не происходит. После произнесения каждого слова и фиксации полученного результата схема автоматически приходит в исходное состояние.

Для проверки надежности действия устройства его выходы были подключены к суммирующей машине, печатающей цифры на бумажной ленте. При произнесении слов *ноль* и *один* на ленте отпечатывались соответственно цифры 0 и 1. Словом *стоп* осуществлялся перевод на следующую строку. В результате испытаний с 10 дикторами обоего пола, каждый из которых произносил по 30 слов голосом и шепотом, на ленте оказались неверно отпечатанными 3 цифры. Надежность действия составила 99%. При произнесении слова *один* пауза перед *дин* не должна была превышать 300 мсек. При нормальном произношении эта пауза (тональная смычка) не превышает 150 мсек. Одна ошибка была вызвана явно неправильным произношением цифры *один* (отсутствовала тональная смычка). Две другие были вызваны погрешностями элементов схемы. Интересно отметить значительные колебания длительности правильно различаемых слов для отдельных дикторов. Продолжительность слова *ноль* была в пределах 200—420 мсек, слова *один* — 240—600 мсек, фонемы *с* в слове *стоп* — 90—170 мсек.

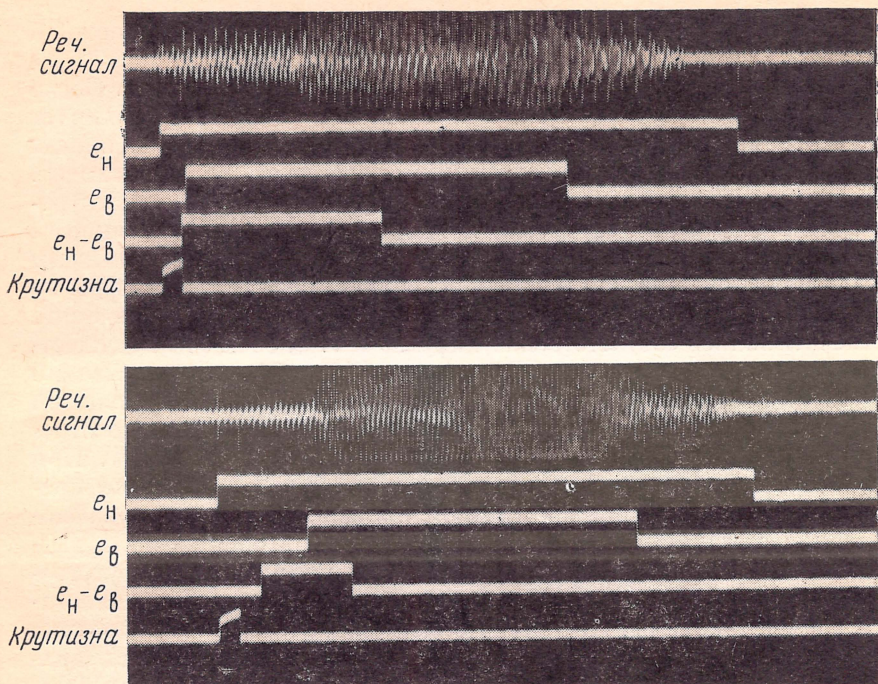
Для увеличения помехоустойчивости устройства можно использовать параметр E_+ и добавить двоичный признак E_1 , пороговое значение которого превышает уровень шума в помещении. При отсутствии E_1 выходные реле не должны срабатывать. Одновременно этим устанавливается минимально допустимая громкость произнесения, что также повышает надежность действия устройства.

Опознавание слов по нескольким групповым признакам

Предварительное исследование проводилось на материале 10 цифр, произнесенных 20 лицами обоего пола. Из речевого сигнала в соответствии с блок-схемой, приведенной на рис. 8, выделялись следующие двоичные признаки [102]:

- 1) наличие энергии в диапазоне до 500 гц — e_n ;
- 2) наличие энергии в диапазоне выше 450 гц — e_b ;
- 4) разность энергии e_n — e_b ;
- 8) усредненная плотность нулей в диапазоне выше 3 кгц — r_t ;
- 16) полная энергия сигнала — E .

В соответствии с сочетаниями этих признаков формировались сегменты слов:



Р и с. 48

Осциллограммы речевого сигнала и групповых признаков для слова *ноль* в мужском (сверху) и женском произнесениях

- 1) звонкий ζ (включая тональную смычку σ') — при наличии только e_n .
- 2) шумный ρ — при наличии только длительного e_v (в случаях $\rho_\tau < 3$ кГц).
- 3) гласный γ — при наличии e_n и e_v .
- 5) назальный ν — при наличии $e_n - e_v$ (и конечно e_n).
- 7) назализованный или низкотоновый гласный при наличии e_n , e_v и их разности.
- 8) шумный ρ — при наличии ρ_τ независимо от других признаков.
- 0) глухая смычка σ — при отсутствии признаков.

Признак E , характеризующий ударные гласные, применялся для переспроса тихо произнесенных слов. Признак ρ_τ более чувствителен, чем e_v , благодаря чему хорошо выделяет окончания слов — *ть*. Гласный и звонкий сегменты длительностью ≤ 25 мсек, а все остальные ≤ 10 мсек не учитывались.

На рис. 48—51 приведены осциллограммы речевого сигнала, признаков e_n , e_v , $e_n - e_v$, а также крутизны нарастания сигнала для слов *ноль*, *один*, *четыре*, *семь* в мужском и женском произне-

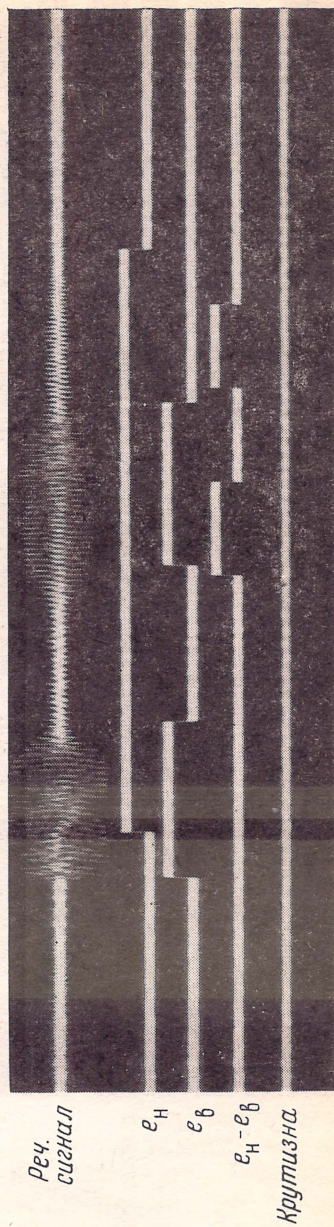
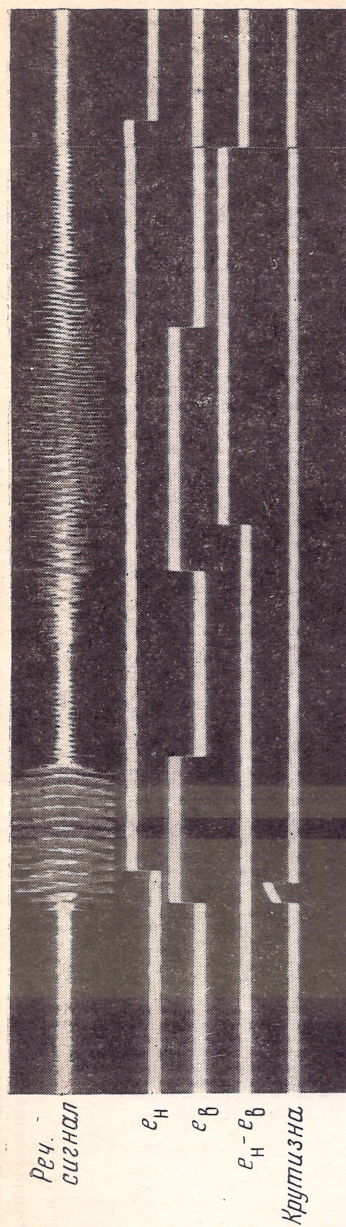
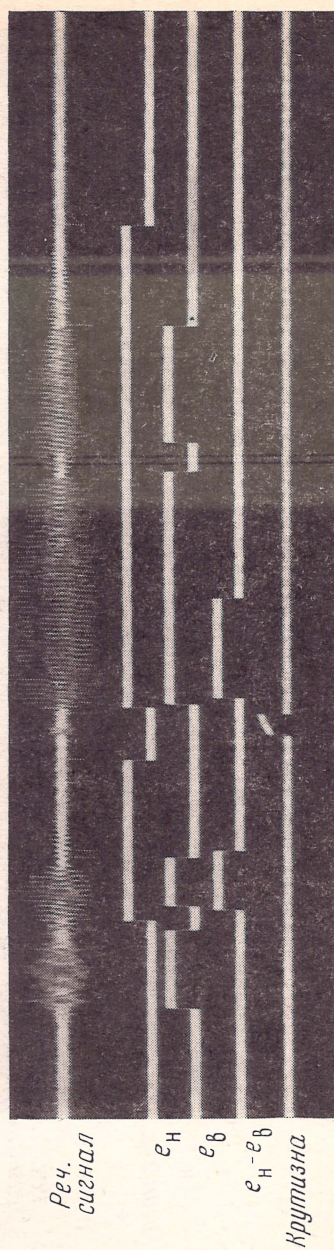
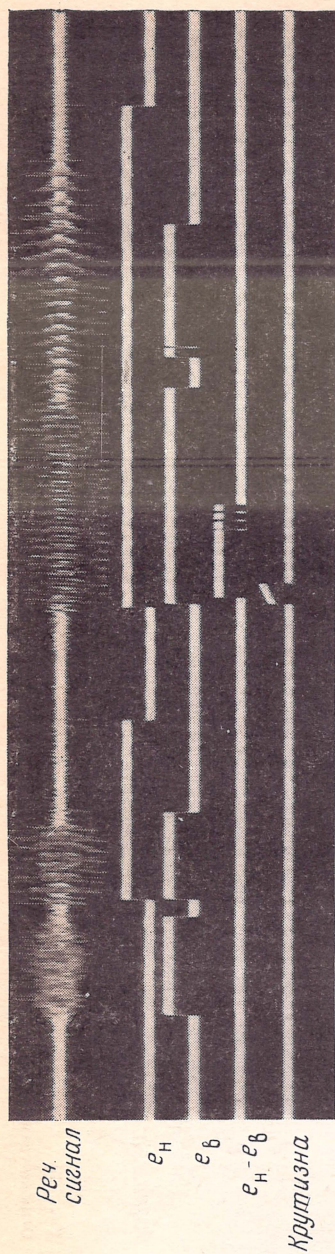


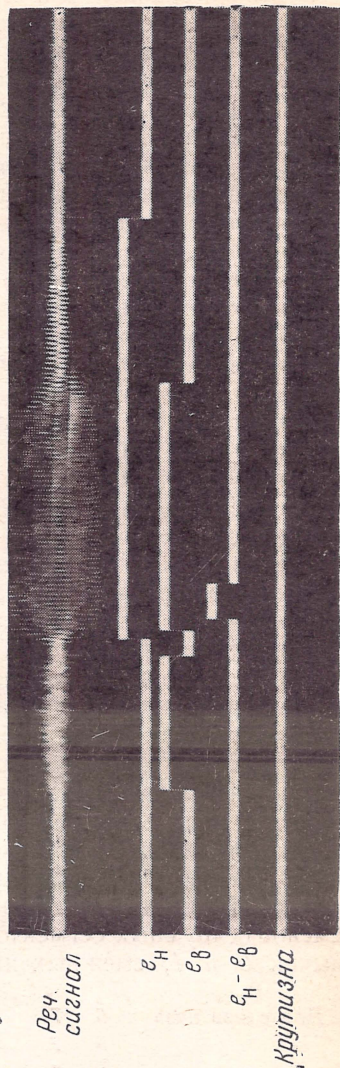
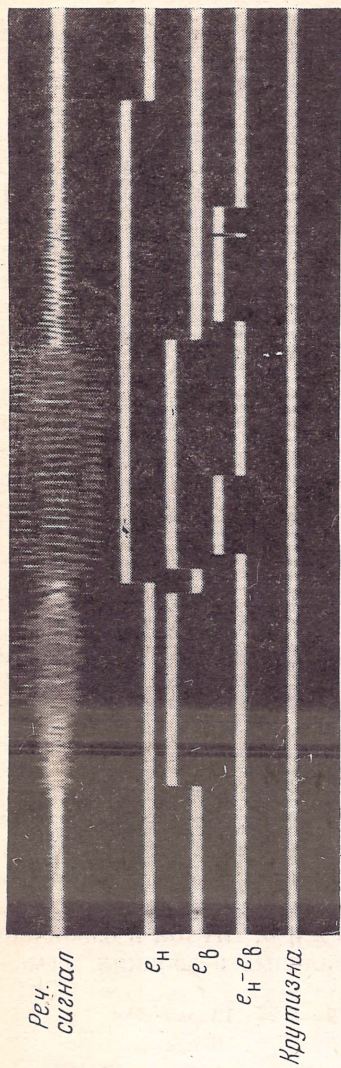
Рис. 49

Осциллограммы речевого сигнала и групповых признаков для слова *один* в мужском (сверху) и женском произнесениях



Р и с. 50

Осциллограммы речевого сигнала и групповых признаков для слова *четыре* в мужском (сверху) и женском произнесениях. Начальная часть глухой смычки перед *т* в обоих случаях озвончена!



Р и с. 51
Осциллограммы речевого сигнала и групповых признаков для слова 'семь' в мужском (сверху) и женском произнесениях

сениях. Пороговые значения признаков устанавливались исходя из речевого процесса, который они должны отражать. Так, с одной стороны, признак e_n должен быть зафиксирован при тональной смычке, где сигнал обладает небольшой энергией. С другой стороны, он не должен срабатывать от шума помещения и не должен перекрывать всю длительность глухой смычки после звонких звуков. Порог $e_n - e_b$ определялся по менее интенсивным носовым звукам (расположенным в конце слов), но при этом он срабатывал и на низкотональные гласные некоторых лиц.

Переходя к анализу полученных последовательностей сегментов слов набора, отметим, что многим одинаковым фонетическим элементам соответствуют определенные сочетания сегментов. Так, например, глухая смычка после звонкового звука отображается сочетанием $1\ 0$ ($\sigma'\sigma$). Начало смычки звонкое, что можно объяснить главным образом реверберацией и также тем, что прекращение колебания голосовых связок происходит в отдельных случаях после смыкания речевого тракта. В граничных случаях один из отрезков отсутствует, т. е. смычка становится или чисто глухой или чисто тональной. Аналогично озвончается начальная часть шумовых сегментов c , $ч$, $ц$ и отчасти $ш$, расположенных после звонких. Однако при небольшом наборе слов можно глухие и звонкие шумные сегменты учитывать вместе. Начальное a характеризуется последовательностью $2\ 3$ ($e_b\gamma$), причем длительность e_b невелика и т. д.

Для одинаковых слов часть сегментов, как по совокупности признаков, так и по порядку следования друг за другом, совпадает. Эти отрезки содержат наиболее существенную информацию, по которой определяется принадлежность данной последовательности к одному из слов набора. Несовпадение остальной части сегментов можно объяснить двумя причинами: 1. Появление нехарактерных, обычных коротких отрезков, большей частью в местах переходов между звуками. Значительную часть этих отрезков можно в данном случае не учитывать. 2. Имеющиеся признаки недостаточно характеризуют определенные звуки и звукосочетания ($в$, $дв$, $дев$).

Ниже приведены варианты последовательностей сегментов (элементов подмножеств C_k) для слов *ноль* и *шесть* с указанием числа реализаций каждого варианта (при объединении эквивалентных смежных шумных сегментов 2 и 8, глухой и озвонченной частей смычек 0 и 1, смежных назальных и звонких сегментов 5 и 1).

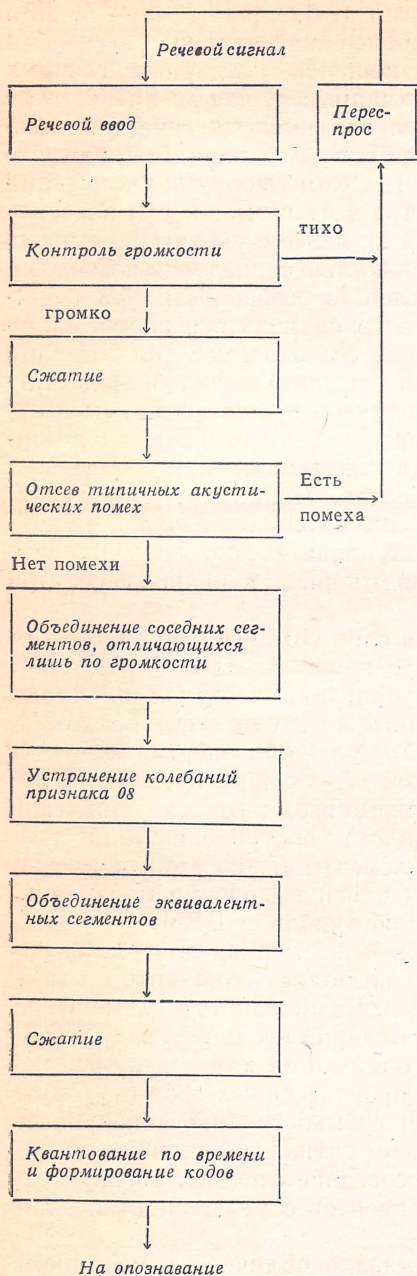
Для слова <i>ноль</i>	5	7	19	1	13 раз
	5	7	19	5	6 раз
	3	7	19	1	1; всего 20.
Для слова <i>шесть</i>	8	19	1	8 0 8	9 раз
	8	19		8 0 8	8
	2	19		8 0 8	2
	2	19	1	8 0 8	1; всего 20.

Число вариантов слова *пять* составило 3, слова *семь* — 5. Для остальных цифр число вариантов более значительно, что отчасти объясняется различиями в произношении неударных гласных вплоть до полного их отсутствия (например, *четыре* или *чтыре* вместо *четыре*). Однако во всех словах имеются сочетания сегментов, одинаковые для всех дикторов. Все варианты каждого слова отличаются от вариантов других слов набора (с учетом длительности некоторых сегментов), что дает возможность опознать все исследованные реализации. Тем не менее с увеличением числа дикторов следует считаться с возможностью совпадения последовательностей сегментов различных слов, в особенности при нечетком произношении, и применять в этих случаях переспрос. Упрощение последовательностей сегментов (необходимое для создания более компактной схемы принятия решения) может быть достигнуто отбрасыванием некоторых признаков в сочетаниях, не несущих существенной информации в общем случае, и для слов данного набора в частности, не считая уже выполненного объединения эквивалентных смежных сегментов. Например, набор слов, состоящий из 10 цифр, содержит носовые только в начале или конце слова. На остальных участках слова можно не учитывать признак назальности и этим уменьшить число вариантов сегментов в средней части слов.

Опознавание небольшого набора слов (10 цифр) по описанным выше признакам производилось с помощью ЦВМ БЭСМ-2М. Речевой сигнал с микрофонного усилителя поступал на блоки выделения признаков, вмонтированные в одну из стоек БЭСМ-2М. Был добавлен еще один признак 20 (разность энергии $e_v - e_n$), имевший по сравнению с признаком $e_n - e_v$ другие весовые коэффициенты. После установки команды «ввод с голоса» появление любого первого признака приводило к поступлению опросных импульсов (с частотой 100 гц) на элементы I , ко вторым входам которых были подключены выходы цепей выделения признаков. С выходов элементов I признаки поступали в ЦВМ.

Для речевого ввода использовалась временная диаграмма читающего устройства БЭСМ-2М, что позволяет сохранить временную диаграмму ЦВМ при минимальных изменениях в схеме ввода (описанном подробнее в [82]). По окончании ввода вступала в действие программа препарирования, блок-схема которой приведена на рис. 52. При обнаружении тихо произнесенных слов (по отсутствию признака E_1) блок контроля громкости приостанавливает дальнейшее действие программы и дает сигнал переспроса. Следующий блок объединял одинаковые соседние отсчеты, образуя таким образом последовательность сегментов с указанием длительности каждого из них.

Появление следующего блока вызвано наличием в машинном зале довольно сильного шума (с давлением в полосе НЧ примерно $0,15 \text{ н}^{-1}\text{м}^2$), что вызывает напряжение в цепи признака e_n ,



Р и с. 52

в 2 раза превосходящее напряжение при тональной смычке для начальных взрывных и в 40 раз — собственные шумы цепи. Наличие сильного шума заставляло поднять порог срабатывания признака e_n . Все же в ряде случаев шум превышал и более грубый порог, что сделало необходимым введение блока отсева акустических помех типа 10, что соответствует появлению с перерывом одного лишь признака e_n .

Далее происходило объединение соседних сегментов, различающихся только по громкости (наличием или отсутствием признака E_T). Блок устранения колебаний объединял в начале слов последовательности сегментов типа 808. Наличие подобных последовательностей вызывалось колебаниями признака 8 по частоте вблизи порогового значения (главным образом для звука *ш* в слове *шесть*). Следующий блок объединял эквивалентные сегменты, а также сегменты, которые могут быть объединены в данном частном случае (набор слов состоит из 10 цифр, имеются сильные акустические помехи). Программа препарирования завершалась квантованием сегментов по длительности и формированием кодов, представлявших последовательности сегментов в окончательном (препарированном) виде. Сегменты разделялись по длительности на 4 градации: 10—50 мсек, 50—100 мсек, 100—200 мсек и более 200 мсек с соответствующими обозначениями 00, 40, 80, 120 (или 20 в 16-ричной системе исчисления,

принятой в ЦВМ БЭСМ-2). Эти числа добавлялись к обозначению сегмента. Например, сегмент 8 в зависимости от длительности записывался как 08, 48, 88 или $\bar{2}8$.

Для иллюстрации действия программы препарирования ниже приведены примеры кодов для слов: *один*, *пять*, *восемь*: 1) до блока объединения сегментов, различающихся только по громкости, и 2) в окончательном виде. В первом случае слева от обозначения сегмента указана его длительность в 10^{-2} сек.

Для слова *один*:

- 1) 1 02, 2 03, 2 13, 4 33, 1 37, 1 05, 8 01, 9 08, 1 05, 8 17, $\bar{2}$ 05, 8 01.
- 2) 02 73 05 41 48 05 53 85

Для слова *пять*:

- 1) 2 05, 3 17, 1 37, $\bar{3}$ 33, 1 13, $\bar{1}$ 01, 1 00, $\bar{3}$ 08.
- 2) 05 37 $\bar{1}3$ 81 00 88

Для слова *восемь*:

- 1) $\bar{0}$ 01, 1 05, 3 17, 8 37, 1 17, 2 05, $\bar{5}$ 08, 2 05, 3 17, 8 05, 0 01.
- 2) 81 05 $\bar{1}3$ 05 88 05 13 45

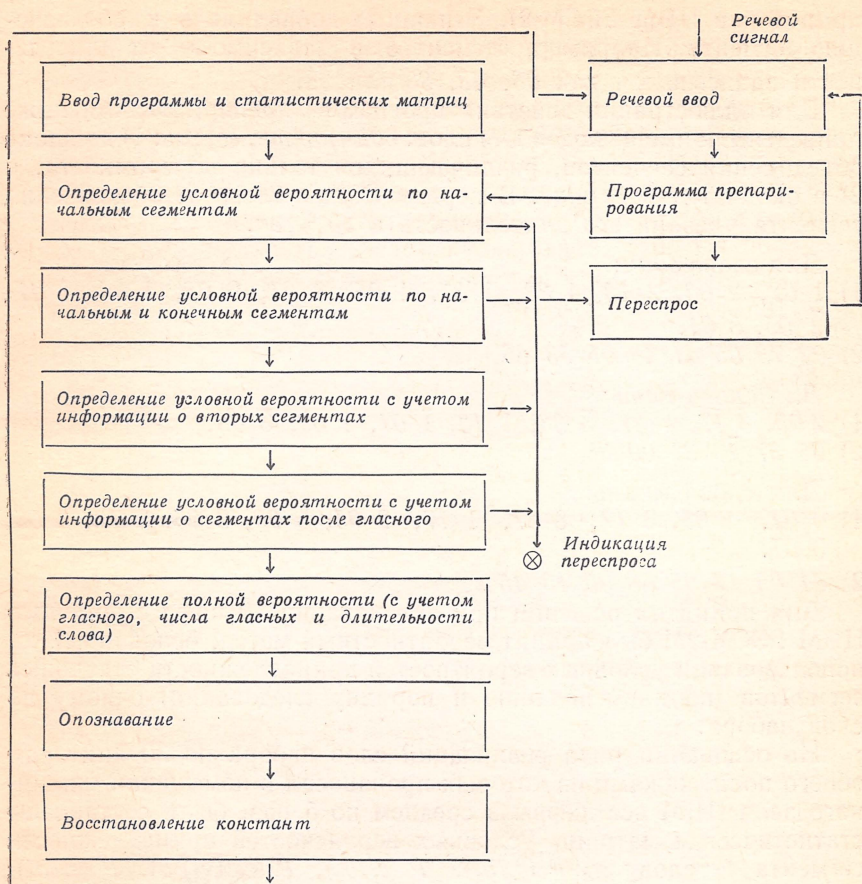
Для принятия решения при опознавании 10 цифр с помощью ЦВМ БЭСМ-2М был принят вероятностный метод, основанный на использовании условных вероятностей принадлежности отдельных сегментов (по наименованию и порядку следования) одному из слов набора.

На основании ряда реализаций слов набора по 30 дикторам обоего пола, каждый из которых произносил в помещении машинного зала ЦВМ все цифры в среднем по 5 раз, были составлены статистические матрицы условных вероятностей принадлежности сегмента b_i слову c_k : $P(c_k/b_1)$, $P(c_k/b_2)$, $P(c_k/(\gamma_1) b)$, $P(c_k/b_z)$, $P(c_k/\gamma_1)$, $P(c_k/n\gamma)$ и $P(c_k/T)$, где c_k — одно из слов набора; b_1 , b_2 , b_z — первый, второй и последний сегменты слова; γ_1 — первый гласный сегмент; $(\gamma_1) b$ — первый сегмент после первого гласного; $n\gamma$ — число гласных; T — длительность слова. Редко встречаемые по положению и наименованию сегменты (менее 0,3% общего числа реализаций) в статистические матрицы не вошли.

Принимая в целях упрощения программы опознавания, что условные вероятности $P(c_k/b_i)$ не обладают взаимной корреляцией, получим полную вероятность

$$P(c_k) = \prod_{i=1}^7 P(c_k/b_i).$$

Фактически некоторая зависимость между статистическими матрицами имеет место.



Р и с. 53

Блок-схема программы опознавания приведена на рис. 53. В первом блоке этой программы происходило сравнение первых сегментов кодов, поступающих после препарирования, с матрицей $P(c_k/b_1)$. При совпадении наименований сегментов произнесенного слова и матрицы из соответствующей строки выбирались значения условных вероятностей для всех 10 слов. Если в матрице не находилось нужного сегмента (в случае редко встречаемой реализации), выдавался сигнал переспроса. Аналогичным образом происходило сравнение поступившего кода с матрицами в последующих блоках. По выбранным из матриц условным вероятностям вычислялись вероятности $P(c_k)$ для всех 10 слов. Наиболее вероятное слово, определяемое по наибольшей $P(c_k)$, выводилось на печать.

Проверка эффективности выбранных признаков пороговых значений, программ препарирования и опознавания производилась по 16 дикторам (9 мужчин, 7 женщин). Каждый диктор произносил все цифры по 2—3 раза. Некоторые из этих дикторов использовались при составлении статистических матриц. Остальные были впервые привлечены к эксперименту. В результате проверки был правильно опознан 91% цифр, случаев переспроса было 1,75%, ошибочно опознано 7,25%. Ошибки вызваны замеченными дефектами произнесения, обнаруженными недостатками программ, а также совпадением в отдельных случаях кодов для слов *ноль* и *два*. При сравнении этих результатов с работами других авторов следует учесть, что в данном случае опознавание производилось в условиях сильного акустического шума.

Результаты исследования показали, что групповые признаки, даже при приближенном выделении из сигнала, позволяют осуществлять опознавание небольшого набора слов. Примененный метод принятия решения показал себя недостаточно эффективным, если сравнить полученную надежность опознавания с объемом необходимых вычислений и требуемой памяти. Описанный эксперимент изложен подробнее в [82, 102].

§ 4. Опознавание нескольких десятков слов

Работа по опознаванию нескольких десятков слов проводилась с целью установления непосредственной связи человека с машиной. ЦВМ должна опознавать слова, произнесенные в микрофон, и выполнять по ним необходимые действия, результаты которых выдаются на печать. Получение устного ответа от ЦВМ при ограниченном словаре не должно вызывать принципиальных трудностей, так как нужные слова могут быть заранее записаны и по мере необходимости воспроизведены.

При опознавании нескольких десятков слов целесообразно иметь две системы признаков — на уровне сегментов (фонем) и на уровне слов. Как уже указывалось, по полученным до настоящего времени признакам может быть опознана лишь часть сегментов, остальные могут быть классифицированы с большей или меньшей полнотой по способу образования звука, что указывает на принадлежность этих сегментов лишь к той или иной группе звуков.

Для опознавания слов использовались признаки, позволяющие определить групповую принадлежность сегментов, и признаки квазистационарных сегментов ударных гласных и щелевых, выделяемые из клиппированного речевого сигнала. Описание этих признаков и методов их получения было изложено в предыдущей главе. По полученным последовательностям сегментов определялись словесные признаки. Решение о принадлежности слова принималось методом логических деревьев с помощью алгоритмов, опи-

санных ниже. При составлении алгоритмов в первую очередь применялись более надежные словесные признаки, характеризующие соотношения отдельных сегментов внутри слова и групповые признаки.

Описание экспериментов

Речевой материал вначале состоял из 41 слова (включая 10 цифр), предназначенных для управления ЦВМ голосом при выполнении на ней некоторых математических операций. Позднее словарь был расширен до 58 слов. Описание алгоритма отражает этап наличия лишь первых 41 слова. Перечень слов и их коды приведены в табл. 17.

Т а б л и ц а 17

Перечень опознаваемых слов и их коды

1. ноль	00	21. косинус	24	41. градус	50
2. один	01	22. запись	25	42. стереть	51
3. два	02	23. ячейка	26	43. степень	52
4. три	03	24. экспонента	27	44. число	53
5. четыре	04	25. целых	30	45. скобка	54
6. пять	05	26. десятых	31	46. целый	55
7. шесть	06	27. точка	32	47. исправить	56
8. семь	07	28. лента	33	48. вычислить	57
9. восемь	10	29. конец	34	49. арктангенс	60
10. девять	11	30. корень	35	50. аргумент	61
11. сложить	12	31. барабан	36	51. интеграл	62
12. умножить	13	32. слушай	37	52. цикл	63
13. вычесть	14	33. плюс	40	53. двоеточие	64
14. разделить	15	34. минус	41	54. функция	65
15. синус	16	35. арккосинус	42	55. массив	66
16. логарифм	17	36. тангенс	43	56. запятая	67
17. пуск	20	37. котангенс	44	57. закрыть	70
18. арксинус	21	38. ввести	45	58. очистить	71
19. печать	22	39. вывести	46		
20. стоп	23	40. сумма	47		

Для получения учебной выборки было привлечено 30 дикторов (20 мужчин, 10 женщин). Слова произносились в помещении машинного зала ЦВМ с относительно высоким уровнем шума (примерно 60 дб), вызванным работой печатающего устройства, магнитных барабанов и вентиляции. Это обстоятельство вынуждало говорить с повышенной громкостью. Речевой сигнал с микрофона МД-57 поступал на устройство выделения признаков сегментов (звуков), соединенное непосредственно с БЭСМ-3М. Схема связи устройства и ЦВМ описана в статье Высоцкого и Трунина-Дон-

ского [22]. В оперативную память ЦВМ каждые 10 мсек записывались значения 4 групповых бинарных признаков (энергии в нижней и верхней областях частот — e_n и e_v , усредненной плотности нулей ρ , полной энергии E) и каждые 20 мсек двух признаков гласных — ΔF_1 и ΔF_2 , по которым могут быть приближенно определены первые две формантные частоты F_1 и F_2 : $F_1 \cong \Delta F_1 / 0,02$; $F_2 \cong \Delta F_2 / 0,02$.

Кроме того, один раз записывались 4 признака каждого длительного шумного звука, указывающие количество интервалов между нулями $2\Delta N$, число интервалов (соответствующих положительным импульсам сигнала) длительностью более 0,4 мсек и более 0,2 мсек — Δn_1 и Δn_2 и число интервалов длительностью менее 0,05 мсек — Δm_1 . У признаков гласных и шумных звуков записывались только 6 старших двоичных разрядов. По окончании произнесения слова в оперативной памяти оказывалась записанной последовательность признаков сегментов во времени.

Образец записи для слова *четыре* приведен в табл. 18.

Квазистационарные сегменты слова показаны фигурными скобками. Слева указано число одинаковых, следующих друг за другом строк. Крайние 2 столбца справа занимают групповые признаки

Таблица 18

Последовательность признаков для слова *четыре*

	ΔF_1	ΔF_2	Δn_1	Δm_1	Δn_2	ΔN											
3	00	00	00	00	00	00	06	}		1	10	07	00	00	00	13	
2	00	00	00	00	00	00	07			1	05	00	00	00	00	00	03
1	00	00	01	22	04	32	07			1	04	00	00	00	00	00	01
1	00	00	00	00	00	00	07			3	00	00	00	00	00	00	01
1	00	00	00	00	00	00	03										
1	05	23	00	00	00	00	03										
1	07	15	00	00	00	00	13										
1	04	00	00	00	00	00	03										
7	00	00	00	00	00	00	00	}									
1	00	00	00	00	00	00	03										
1	04	16	00	00	00	00	03										
1	06	24	00	00	00	00	13	}									
2	05	21	00	00	00	00	13										
1	05	22	00	00	00	00	13										
1	05	21	00	00	00	00	13										
1	05	22	00	00	00	00	13										
1	05	21	00	00	00	00	13										
1	05	23	00	00	00	00	13										
1	07	21	00	00	00	00	03										
1	05	22	00	00	00	00	13										
1	10	20	00	00	00	00	13										
1	15	21	00	00	00	00	13										
1	14	21	00	00	00	00	13										

e_n — 01, e_b — 02, p — 04 и E — 10. Каждый из остальных признаков выражается двузначным восьмеричным числом и занимает два столбца. Крайние 4 столбца слева — признаки гласных ΔF_1 и ΔF_2 (последний без младшего разряда). На остальных столбцах расположены признаки длительных шумных звуков Δn_1 , Δm_1 , Δn_2 и ΔN (считая слева направо). У признаков Δm_1 и ΔN отбрасываются соответственно 2 и 3 младших разряда.

Наличие признака p в первых семи строках указывает, что слово *четыре* начинается с длительного шумного звука. В соответствии с признаками шумных (см. § 5 гл. 3) это или *ш* или *ч*. Появление признака 01, начиная с середины *ч*, характеризует озвончение его конечной части. Далее следует короткий гласный *и* (по данным F_1 и F_2 , см. § 3 гл. 3), соответствующий редуцированному неударному *'э*, обычно произносимому в этом слове как *и*. После гласного располагается глухая смычка, указывающая на наличие глухого взрывного звука. Ввиду малой длительности шумового сегмента *т* p отсутствует. Затем следует длительный машинный гласный, соответствующий звуко сочетанию *ыре*. Начало его *и*-образное (*и* и *ы* по признакам, описанным в § 3 гл. 3, входят в одну группу), ксонец *э*-образный, завершающийся низкочастотным «гулом» длительностью в 40 мсек.

Образцы записи для слова *стоп*, произнесенного в нормальном и более быстром темпе, приведены в табл. 19.

Т а б л и ц а 19

Последовательность признаков для слова *стоп*

1) 1 00 00 00 00 00 00 04	2) 1 00 00 00 00 00 00 04
1 00 00 00 00 00 00 06	1 00 00 00 00 00 00 06
1 00 00 00 75 00 60 06	1 00 00 00 44 00 42 06
2 00 00 00 00 00 00 06	1 00 00 00 00 00 00 06
4 00 00 00 00 00 00 00	1 00 00 00 00 00 00 04
1 00 00 00 00 00 00 03	3 00 00 00 00 00 00 00
1 04 12 00 00 00 00 03	1 02 02 00 00 00 00 03
1 14 13 00 00 00 00 13	1 12 13 00 00 00 00 13
1 14 06 00 00 00 00 13	1 12 14 00 00 00 00 13
1 13 07 00 00 00 00 13	1 12 15 00 00 00 00 13
1 13 13 00 00 00 00 13	1 12 14 00 00 00 00 13
1 13 06 00 00 00 00 13	1 10 05 00 00 00 00 13
1 04 00 00 00 00 00 01	1 00 00 00 00 00 00 01
1 00 00 00 00 00 00 01	3 00 00 00 00 00 00 00
3 00 00 00 00 00 00 00	3 00 00 00 00 00 00 01
1 00 00 00 00 00 00 02	
1 00 00 00 00 00 00 03	
1 06 00 00 00 00 00 03	
1 04 00 00 00 00 00 03	
1 00 00 00 00 00 00 03	
2 00 00 00 00 00 00 01	

Коды слова, введенные в МОЗУ, могут занимать до 100 ячеек памяти. В связи с тем, что отсчеты групповых признаков производятся каждые 10 мсек, а два признака гласных ΔF_1 и ΔF_2 снимаются через каждые 20 мсек, для ускорения работы опознающего алгоритма полезно иметь трансформированное (сжатое) поле последовательностей сегментов, которое занимало бы примерно в два раза меньше ячеек, чем первоначальное поле признаков. При формировании сжатого поля исключаются отсчеты, расположенные между строками, содержащими значения ΔF_1 и ΔF_2 , с сохранением информации о признаках шумных звуков. Такое поле формируется перед работой программы опознавания. Его полезно иметь и потому, что при выдаче информации о последовательностях признаков слова на печать время печати сокращается почти вдвое, а существенной потери информации не происходит. Следует отметить, что первоначальное поле продолжает храниться в памяти и некоторые блоки опознающего алгоритма работают, анализируя первичное поле.

Формировалось еще третье поле — последовательности машинных гласных слова. Необходимость в его создании объяснялась также стремлением ускорить работу некоторых блоков алгоритма. В этом поле, построенном на основе анализа F_1 и F_2 , указывается код соответствующего гласного участка и его длительность. Последовательность машинных гласных слова *ноль* может, например, быть представлена таким образом: 5н 1у 4о 3н, что означает временную последовательность: 5 участков н, 1 участок у, 4 участка о, где длительность каждого участка — 20 мсек. Отметим, например, что блок длительности слова T опознающего алгоритма работает с первичным полем, блок поиска формантных частот — со сжатым, а блоки гласных а, о, э, и, у — с полем последовательностей машинных гласных.

Алгоритм опознавания

В результате статистической обработки учебной выборки были получены признаки слов набора. По трем из них — наличию длительного шумного в начале, середине и конце последовательности — ρ_n , ρ_c , ρ_k был составлен алгоритм для разделения на ЦВМ всех слов набора на 8 групп, причем некоторые слова в зависимости от особенностей и вариантов произношения попадали в две и более групп. Рассмотрим, например, разделение слов по признаку ρ_n . Для слов *девять* и *десятых* $\rho_n = 1$ лишь при'affрицированном произношении и сравнительно коротком сегменте тональной смычки ∂' , что имело место в 22,9% случаев. В остальных случаях $\rho_n = 0$, так как шумовой сегмент ∂' был непродолжителен или располагался после длительной тональной смычки. В словах *восемь* и *вычесть* $\rho_n = 1$ соответственно в 6 и 2,8% случаев при наличии сильной шумовой составляющей *в*. В слове *запись* $\rho_n = 0$ в 15,6%

случаев при длительном голосовом сегменте в начале з. Из числа остальных 36 слов набора для 9 всегда $\rho_n = 1$, для 27 слов $\rho_n = 0$.

Блок-схема логического дерева членения слов набора на группы приведена на рис. 54.

При разделении слов набора на 8 групп учитывались все варианты произношения. Для каждой из этих групп был составлен отдельный алгоритм опознавания в виде логического дерева, использующий наиболее информативные признаки слов группы. Большинство словесных признаков нашло применение в нескольких алгоритмах, в том числе такие, как длительность слова, число сегментов с наличием признаков гласных, наличие отдельных щелевых и гласных в определенных местах слова, соотношения по длительности определенных сегментов слов и др. Каждой группе слов присвоен код, приведенный на рис. 54 внизу.

Рассмотрим подробно в качестве примера сравнительно сложный алгоритм опознавания слов группы 06. В эту группу входят реализации слов *девять*, *умножить*, *вычест*, *разделить*, *пуск*, *арксинус*, *печать*, *косинус*, *запись*, *десятых*, *арккосинус*, имеющие длительные шумные звуки (сегменты) в середине и конце слова. Из общего числа 11 слов в одном может быть только 1 «машинный» гласный, в двух — 1—2, в трех — 2—3 и в двух — 3 «машинных» гласных. В большинстве слов в случае наличия двух-трех «машинных» гласных первый из них менее информативен, так как он либо является неударным, либо образуется звонким согласным. Отсюда следует 2 вывода: в первую очередь необходимо выделить реализации слов с одним «машинным» гласным, которые составляют около 10% общего числа реализаций группы. Для остальных слов отсчет гласных лучше вести от конца слова. Наиболее информативным признаком группы является отношение длительностей предпоследнего и последнего гласных τ_{y_1} и τ_{y_2} .

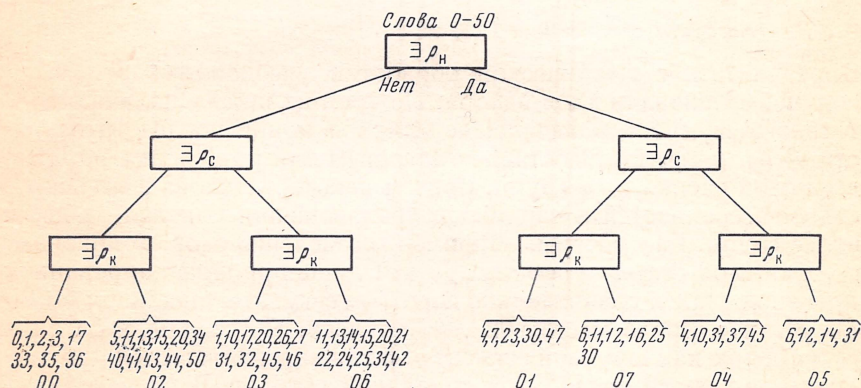
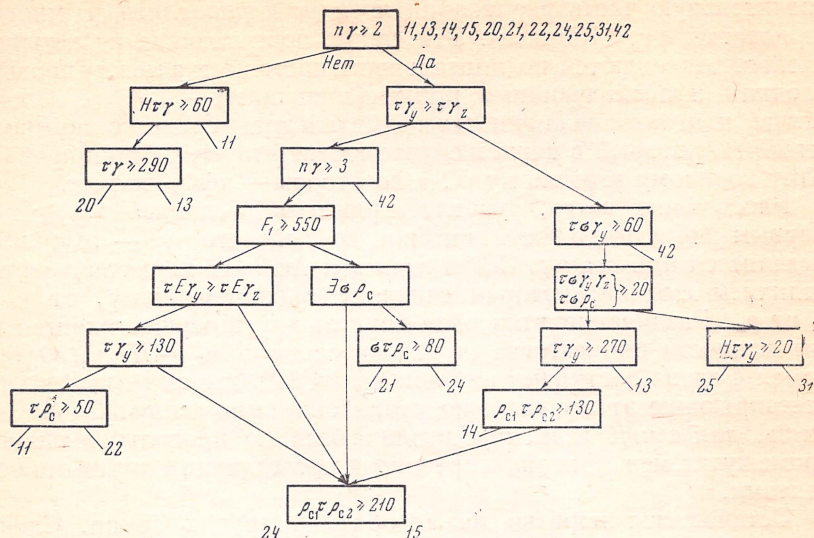


Рис. 54



Р и с. 55

Блок-схема алгоритма опознавания слов группы 06 приведена на рис. 55. В начале логического дерева производится классификация по признаку $ny \geq 2$. В левую ветвь попадают все слова *пуск* и часть реализаций *девятъ* и *умножить*, которые хорошо различаются по признакам Hty (длительность промежутка между началом слова и гласным) и ty (длительность гласного). В основной правой ветви первым стоит блок $ty \geq tyz$. Бóльшая длительность y_u указывает также на более громкое произнесение этого гласного (в особенности, если он является ударным) и на возникновение в связи с этим в части реализации слов (например, *вы-честь*, *запись*, *разделить*) дополнительных предшествующих y_u «машинных» гласных. При $ty < tyz$ дополнительных «машинных» гласных не возникает, и все указанные выше слова представлены реализациями, содержащими только 2 гласных. Это дает возможность сразу выделить слово *арккосинус* рассматриваемой ветви как единственное, содержащее 3 и более гласных. Из числа оставшихся пяти слов отделяется затем *арксинус*, как удовлетворяющий условиям: $F_1 y_u \geq 550$ (звук *a* произносится в нем достаточно четко) и $\sigma tpc < 80$ (длительность промежутка от смычки до среднего шумного звука меньше 80 мсек). У редких реализаций *косинус*, попадающих в этот блок при наличии «машинных» смычек в начале слова, $\sigma tpc \geq 80$. Далее по отношению длительностей огибающих (энергии) предпоследнего и последнего гласных $tey_u \geq tey_z$ и длительности ty_u выделяются подгруппы слов: *де-вятъ*, *печать* и *косинус*, *разделить*. Блок $tey_u \geq tey_z$ выделяет

значительную часть реализаций *косинус* и *разделить*, у многих из которых $\tau_{\gamma_y} < 130$ мсек. Эти слова, поступающие из четырех блоков, разделяются по длительности промежутка между первым средним и последующими шумными звуками — $\rho_{c1}\tau_{\rho_{c2}}$. Слова *девять* и *печать* различаются по длительности среднего шумного сегмента, который в первом случае соответствует аффрицированному звонкому взрывному d' , а во втором — аффрикату $ч$.

Рассмотрим ветвь, удовлетворяющую условию $\tau_{\gamma_y} \geq \tau_{\gamma_z}$. Первым по длительности смычки до гласного γ_y — $\tau\sigma_{\gamma_y} \geq 60$ выделяется *арккосинус*. Слова *запись* и *десять* характеризуются наличием смычки, которая отделяет одновременно γ_y от γ_z и ρ_0 от ρ_k . Различение этих слов друг от друга осуществляется по длительности промежутка от начала слова до γ_y — $H\tau_z\gamma_y$. Оставшиеся 4 слова поступают на блок $\tau_{\gamma_y} \geq 270$, где выделяются удовлетворяющие этому условию реализации *умножить*. Слова *вычесть*, *разделить* и *косинус* различаются по продолжительности промежутка между первым средним и последующим шумными сегментами.

Остановимся коротко на алгоритмах других групп. Слова, в которых нет длительных шумных звуков, образуют группу 00. Вероятность попадания в эту группу слов *ноль*, *два*, *три*, *корень*, *лента*, *барабан* близка к 1, слова *логарифм* — к 0,3 (при нечетком произнесении ϕ), слова *один* — к 0,25 (при слабом аффрицировании d').

Классификация производится вначале по длительности T , чем разделяются в основном односложные и трехсложные слова, затем по числу гласных n_{γ} . Большую роль в этой группе играют словесные признаки, использующие F_1 и F_2 , и наличие определенных гласных.

К числу более сложных можно отнести также алгоритмы групп 02 и 03. По первому из них 11 слов вначале разделяются по длительности промежутка от начала слова до конечного шумного сегмента $H\tau_{\rho_k}$, к которому в целях лучшего разделения слов *градус*, *тангенс*, *девять*, *пять* на длинные и короткие добавляется разность длительностей $\tau\sigma' - \tau\sigma$. Слово *конец* различается от *пять* по наличию провала с минимальным значением не более 350 $\mu\text{с}$ и перепадом F_1 не менее 50 $\mu\text{с}$. В группу 03 входят реализации 10 слов. Наиболее информативным признаком этой группы является наличие (или отсутствие) шумного сегмента до предпоследнего гласного $\tau_{\rho_{\gamma_y}}$. В остальные 4 группы входят от 4 до 6 слов, в связи с чем их алгоритмы получились более простыми. Подробное описание алгоритмов всех групп приведено в [18].

Любое произнесенное слово классифицируется как одно из слов набора. Однако слова, не входящие в набор, располагаются в ветвях логического дерева довольно компактно и опознаются каждое как одно из двух-трех слов набора. Это обстоятельство позволяет считать описанный алгоритм до некоторой степени универсальным,

пригодным без существенных изменений в основных ветвях как для увеличенного набора слов, так и для опознавания других слов, что подтвердилось при добавлении 17 слов к первоначальному 41.

О структуре блоков логических деревьев

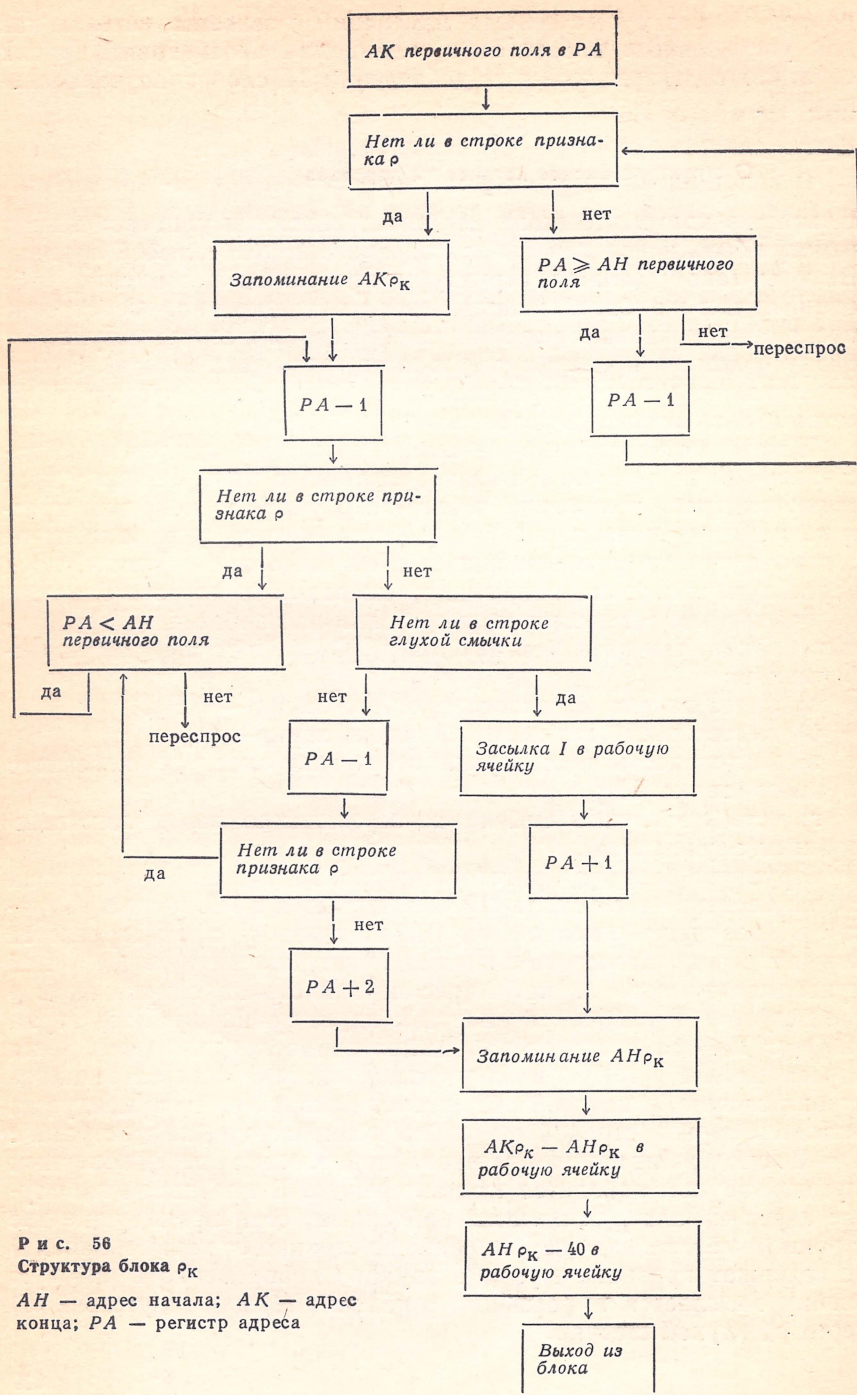
Блоки, из которых состоит алгоритм, являются универсальными и могут быть использованы для опознавания слов любой группы. Выгоды такой блочной структуры программы несомненны, ибо один и тот же блок используется неоднократно. Собственно программа опознавания состоит из обращений к соответствующему блоку и команд проверки результатов, выданных блоком. Значение признака 0 или 1 управляет дальнейшим ходом программы по ветвям логического дерева.

Некоторые блоки дают не только ответ «да» или «нет» на поставленный вопрос, но и выдают различную информацию о длительности, местоположении какого-либо участка слова или признака и другие параметры. Такие блоки, как правило, входят не только в программу опознавания слов какой-либо группы, но и используются как составная часть какого-либо другого блока.

Рассмотрим в качестве примера один из наиболее универсальных блоков, так называемый блок p_k . Он выдает следующую информацию:

- а) адрес начала конечного шумного звука — $АН_{p_k}$,
- б) длительность конечного шумного звука — $т_{p_k}$,
- в) длительность промежутка от начала слова до конечного шумного звука $Нт_{p_k}$;
- г) есть ли глухая смычка непосредственно перед конечным шумным звуком $Ш_{p_k}$.

Структура блока p_k приведена на рис. 56. Вначале адрес конца слова, введенного в память ЦВМ, засылается в РА (регистр адреса). Далее путем уменьшения РА на единицу просматривается каждая строка последовательности признаков на наличие в ней признака шумного звука. Адрес первой строки с признаком p фиксируется в рабочей ячейке как адрес конца шумного звука ($АК_{p_k}$). Подобные же операции производятся для нахождения адреса начала p_k . При этом учитывается, что если встретилась строка отсчета, в которой отсутствует признак p , то предыдущая является первой строкой p_k только в том случае, когда строка с $p = 0$ не содержит групповых признаков, т. е. если перед конечным шумным звуком находится глухая смычка. В противном же случае адрес начала p_k считается найденным, если предыдущие две строки (имеется в виду перемещение по строкам от конца к началу) не содержат признака шумного звука. Одновременно в рабочую ячейку засылается информация о наличии (1) или отсутствии (0) перед p_k глухой смычки.



Длительность шумного звука определяется как

$$\tau_{\rho_k} = (AK_{\rho_k} - AN_{\rho_k} + 1) \cdot 10 \text{ мсек},$$

а длительность промежутка от начала слова до начала конечного шумного как

$$H \tau_{\rho_k} = (AN_{\rho_k} - AN + 1) \cdot 10 \text{ мсек}.$$

В конце блока стоит сформированная при обращении к блоку команда возврата на ветвь программы, использующую полученную информацию.

В результате проверки алгоритмов опознавания слов на материале 20 дикторов, из которых 10 участвовали в составлении учебной выборки, было правильно опознано 93% слов. Время опознавания после окончания произнесения слова составило в среднем 0,5 сек. Правильность произношения не контролировалась. Отклонения от правильного произношения обнаруживались при рассмотрении выведенных на печать последовательностей сегментов. Ошибки при опознавании вызывались следующими причинами: неправильным произношением ряда слов (в том числе пропуском отдельных звуков); запуском временной диаграммы ввода шумами зала, а не началом слова; недостатками алгоритмов опознавания, обнаруженными в процессе анализа полученных реализаций слов.

Метод принятия решения с помощью логических деревьев показал себя перспективным для применения при опознавании нескольких десятков слов. Исследования по использованию описанного алгоритма для опознавания фраз и выполнения по ним необходимых действий дали положительные результаты (см. след. главу). Эксперименты по опознаванию нескольких десятков слов проводились автором совместно с Высоцким, Рудным и Труным-Донским.

В порядке подготовки к опознаванию нескольких сотен слов проверялась возможная надежность признаков переходных сегментов гласных на словах, различающихся лишь местом образования или смягчением одного из согласных. Речевой материал состоял из 40 слов типа: *точка, кочка, бочка, дочка, скучать, стучать, забор, задор, быть, бить* и других. Слова произносились 20 лицами в помещении машинного зала. Выделяемые из речевого сигнала в аналоговом устройстве признаки звонкости ζ , гулкости, шумности ρ , гласности γ , числа нулей N , формантных частот F_1 и F_2 вводились каждые 20 мсек в ЦВМ «Минск-22». Для принятия решения о принадлежности слова использовались метод логического дерева и словесные признаки из числа описанных выше. Разделение на конечных ветвях дерева осуществлялось по взрывным и гласным, опознанным по последовательности сегментов (см. § 7 гл. 3). Решение принималось в реальном масштабе времени. Средняя надежность опознавания слов составила около 85%. Эти исследования проводились автором совместно с Григоряном и Криновым [29].

ОПОЗНАВАНИЕ ТИПОВЫХ ФРАЗ

§ 1. Об одном методе использования лингвистической информации

При построении системы речевой связи человека с машиной необходимо в первую очередь осуществить автоматическое опознавание фраз-заданий. Рассмотрение фразы как последовательности некоторого числа слов без учета смысловых связей между ними приводит к пониженной надежности опознавания фраз. Возникает необходимость учета этих смысловых связей. Однако введение в систему опознавания всей лингвистической информации о данном языке — задача весьма трудная и громоздкая, и навряд ли она может быть решена приемлемым образом в ближайшее время. Решение подобной задачи может осуществляться постепенно, с использованием вначале отдельных лингвистических закономерностей для ограниченного числа типовых фраз.

Рассмотрим метод использования лингвистической информации, позволяющий значительно повысить надежность опознавания типовых фраз [84].

Слова, необходимые для составления фраз-заданий, разбивают на следующие классы в зависимости от их смыслового значения.

1. Слова, определяющие характер задания (тип фразы): *вычислить, исправить, запись, печать* и т. п. При выполнении, например, математических операций могут быть образованы следующие дальнейшие классы слов:

2. Слова, над которыми производятся действия: цифры от нуля до девяти, *плюс, минус, скобка, точка* (для отделения дробной части числа от целой).

3. Слова, с помощью которых выполняются действия над одним числом или выражением в скобках, *синус* и другие тригонометрические функции, *экспонента, логарифм* (натуральный) и т. п.

4. Слова, с помощью которых выполняются действия над двумя и более числами или выражениями в скобках: *сложить, умножить, вычесть, разделить, возвести* (в степень) и т. п.

5. Слова, с помощью которых выполняются более сложные вычисления: *интеграл, пределы* и т. п.

6. Служебные слова: *слушай, число, стоп, конец* и т. п.

Такое разбиение слов на классы дает возможность автоматически определять структуру фразы (порядок чередования отдельных классов слов) по правилам построения фраз, близким к правилам естественной речи, и по части опознанных слов (примерно так, как это имеет место при восприятии фраз человеком). Установление структуры фразы позволяет исправлять ошибки в случаях, когда некоторое слово фразы опознается как близкое к нему в пространстве признаков слово другого класса. Таким образом, становится возможным, исходя из смысла фразы, исправить большинство неправильно опознанных слов. Используя принцип разбиения слов на классы, можно построить язык для речевой связи человека с машиной с помощью ограниченного числа типов фраз-заданий.

Правила составления фраз могут быть (для того же примера выполнения математических операций) примерно такими:

а) Фраза начинается со специального служебного слова, включающего машину на опознавание фразы-задания. Второе слово фразы принадлежит к первому классу и указывает тип задания.

б) Далее всегда следует слово третьего класса, указывающее наименование функции. Следующее слово принадлежит второму классу. Оно может быть начальной цифрой аргумента или, если выражение представляет собой сложную функцию, словом, разделяющим названия двух функций.

Конец аргумента определяется по следующему за ним служебному слову, не пересекающемуся в пространстве словесных признаков со словами аргумента. При выполнении действий, связанных с вычислением тригонометрических функций, после числа (или выражения в скобках) следует слово *градус* или *радиан*. Эти два слова образуют отдельный класс.

После слова окончания аргумента могут следовать слова четвертого класса (при действиях с двумя и более числами) или служебное слово окончания фразы.

в) В случае действий с двумя и более числами (или выражениями в скобках) построение остальной части фразы производится в соответствии с п. б).

Устройство для речевого ввода данных в вычислительную машину, использующее описанный выше метод, может быть реализовано по блок-схеме, приведенной на рис. 57. Речевой сигнал, соответствующий одному из опознаваемых слов, поступает на блок выделения признаков слов ВП. Основываясь на полученных признаках, блоки опознавания слов ОС 1—5 определяют принадлежность произнесенного слова. Блок ОС 1 опознает слова 1 класса, определяющие тип фразы. Блоки ОС 2 и ОС 3 опознают соответственно слова 2 класса (*ноль, один* и др. — при заданиях на вычисление математических выражений) и слова 3 класса (*корень, логарифм* и др.). Блоки ОС 4 и ОС 5 опознают начальное слово фразы (*слушай*) и конечное слово фразы (*конец*). Сигналы, соответству-

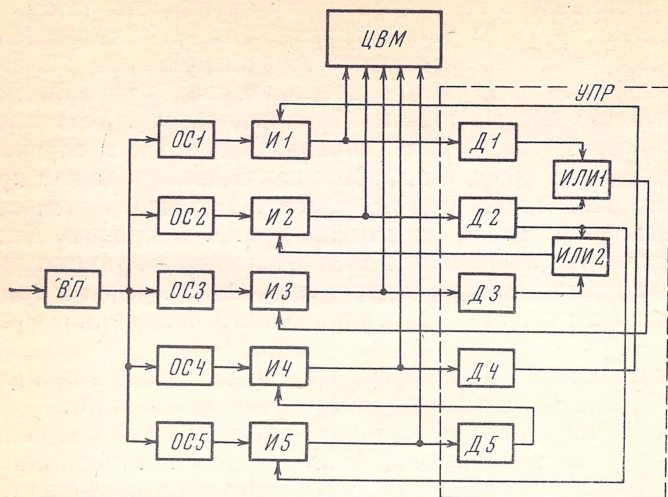


Рис. 57

Блок-схема устройства для речевого ввода данных в вычислительную машину

ВП — блок выделения признаков речевого сигнала; ОС — блок опознавания класса слов; УПР — блок управления; Д — дешифратор; И, ИЛИ — логические элементы

ющие опознанным словам, с блоков ОС 1—5 проходят через логические элементы И 1—5 на вход блока управления, отпирающего или запирающего элементы И 1—5, и на вычислительную машину. В начальном положении блок управления держит открытым элемент И 4 и закрытыми остальные элементы И. В зависимости от наименования опознанного слова или класса слов блок управления запирает часть элементов И, что уменьшает число возможных ошибок опознавания за счет исключения слов, сигналы которых поступают на запертые элементы. Блок управления состоит из дешифраторов Д 1—5 и логических элементов ИЛИ 1, ИЛИ 2.

Рассмотрим действие описанного устройства на примере ввода устного задания на вычисление $\sqrt[3]{\ln 1.357}$, выражаемом фразой: «Слушай вычислить корень скобка логарифм один точка три пять семь конец». Из числа классов слов, опознаваемых блоками ОС, приведенными на рис. 57, второе слово фразы может принадлежать только к 1 классу. Поэтому после произнесения и опознавания блоком ОС 4 первого слова *слушай* и прохождения сигнала через открытый элемент И 4 на блок управления и вычислительную машину дешифратор Д 4 отпирает элемент И 1. При этом дешифратор Д 5, державший открытым И 4, устанавливается в положение, при котором И 4 закрыт. По поступлении следующего сигнала, соответствующего опознанному блоком ОС 1 второму слову *вычислить*, дешифратор Д 1 через элемент ИЛИ 1 отпирает И 3, а И 1 при этом

закрывается, так как после слов 1 класса всегда следуют слова 3 класса. При произнесении каждого из остальных слов фразы блок управления производит аналогичные переключения в соответствии с правилами построения фраз. После слова *корень* блок управления запирает элемент И3 и отпирает И2, после слова *скобка* запирает И2 и отпирает И3, после слова *логарифм* — запирает элемент И3 и отпирает И2. При произнесении первой цифры *один* и каждой из последующих цифр, в том числе слова *точка*, отделяющего дробную часть, блок управления, оставляя открытым элемент И2, отпирает также И5, так как существует возможность того, что любая цифра является последним знаком числа. После произнесения слова *конец* вычислительная машина производит вычисления в соответствии с полученным заданием. Исходные данные задания и результаты действия вычислительной машины выводятся в печать. Одновременно блок управления отпирает элемент И4 и запирает все остальные элементы И, подготавливая таким образом цепи к приему следующей фразы.

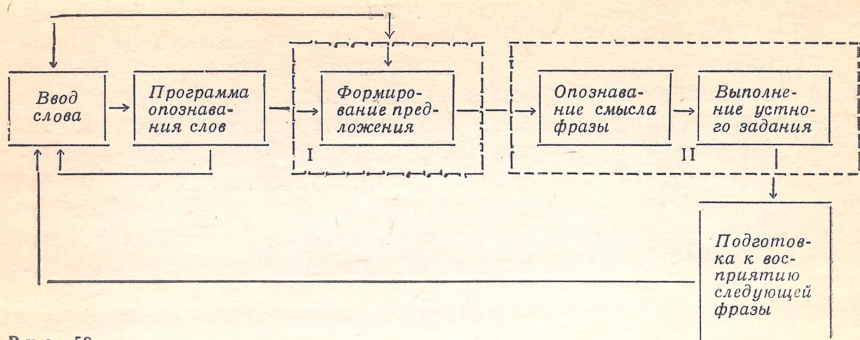
В случаях неверного опознавания фразы оператор повторяет ее. При недостаточной надежности опознавания слов и длинных фразах можно проверять правильность ввода исходных данных в машину по части фразы. Контроль за правильностью опознавания вводимых слов можно осуществлять также визуально с помощью системы отображения информации на электронно-лучевой трубке. Функции блоков опознавания слов, логических элементов И и блока управления могут быть выполнены вычислительной машиной.

§ 2. Опознавание фраз при речевом управлении вычислительной машиной

На основе метода, описанного в предыдущем параграфе, и алгоритма опознавания 58 слов (см. § 4 гл. 4, а также [18]) был создан алгоритм, который позволил формировать из последовательности опознаваемых слов предложения, содержащие различные задания.

Была предусмотрена возможность замены отдельных операций по управлению работой вычислительной машины, выполняемых оператором на пульте (например, внесение исправлений в программу с пульта, контроль содержимого ячеек памяти). Кроме того, можно давать устные задания на выполнение операций обмена между накопителями в вычислительной машине, включая обращение к магнитным барабанам (МБ) и магнитным лентам (МЛ), с использованием как режима записи, так и режима считывания; осуществлять печать (восьмеричную или десятичную) как отдельных кодов, так и массивов.

Алгоритм предусматривает также выполнение заданий на вычисление математических выражений, включая действия над сложными функциями. Так как опознавание слов и формирование



Р и с. 58

предложения происходит в реальном масштабе времени, а при вычислении функций используются стандартные программы, то окончательный результат выдается на печать практически сразу после произнесения последнего слова фразы.

Алгоритм опознавания фраз приведен на рис. 58. Он включает в себя программу опознавания слов и состоит из двух частей: первая часть формирует фразу, а вторая производит ее анализ и выполняет задание. Переход к формированию фразы осуществляется по опознаванию парольного слова *слушай*. Если слово *слушай* не было произнесено, формирования предложения не происходит. Первая часть алгоритма — формирование фраз — состоит из нескольких ветвей (рис. 59), обращение к каждой из которых зависит от типа задания и определяется вторым словом фразы. Дальнейшее описание алгоритма выполняется на примере наиболее сложной ветви — вычисления математических выражений.

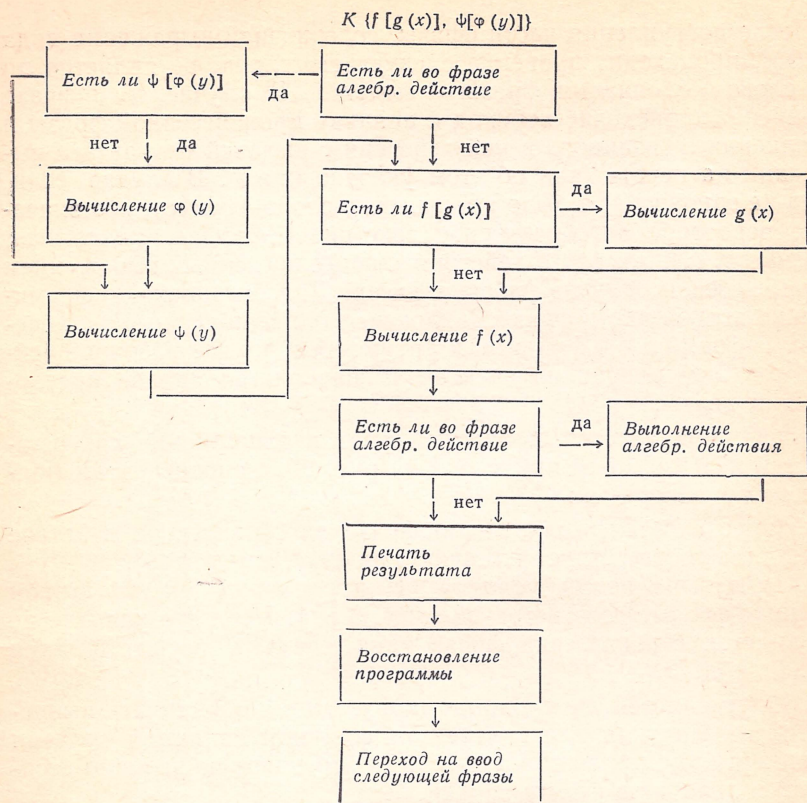
Выполнение заданий на вычисление математических выражений

Рассмотрим более подробно ветвь алгоритма опознавания фраз, содержащих задания на вычисление математических выражений типа

$$K\{f[g(x)], \psi[\varphi(y)]\},$$

где K — алгебраическое действие над функциями f, g, ψ, φ . Аргументы x и y могут быть как положительными, так и отрицательными числами, содержащими до девяти цифр включительно. Для тригонометрических функций аргумент может задаваться как в градусах, так и в радианах.

Для перехода в исходное состояние программы опознавания фраз до окончания произнесения фразы следует сказать служебное слово *очистить*. Любое слово, анализируемое программой формирования фраз (рис. 59), проходит поэтому проверке, не



Р и с. 60

является ли оно словом *очистить*. При поступлении этого слова программа переходит в исходное состояние.

Как уже указывалось, первыми словами фразы должны быть: *слушай*, *вычислить*. При произнесении следующего за ними слова (названия функции или слова *число*) оно опознается, после чего соответствующий код засылается в одну из рабочих ячеек. В другой ячейке устанавливается единица — признак того, что далее должно идти некоторое число (аргумент функции) или слово *скобка*, если вычисляется функция от функции. Формирование аргумента осуществляется с учетом того, что числа могут быть положительными или отрицательными, целыми или дробными. Порядок аргумента определяется числом значащих цифр до слова *точка*, если порядок положительный, и числом нулей после слова *точка* до первой цифры, отличной от нуля, если порядок отрицательный. В случае тригонометрических функций после служебного слова окончания аргумента (*stop*) устанавливается, в каких единицах задан аргумент (в градусах или в радианах).

После поступления части фразы, содержащей выражение вида $f[g(x)]$, происходит проверка следующего слова, не является ли оно словом окончания фразы — *вывести*. В случае, если фраза закончена, происходит переход к анализу произнесенной фразы и выполнению указанного в ней действия с выдачей результата вычислений на печать (как об этом указано ниже). В случае если фраза не закончена, а было произнесено слово 4 класса, обозначающее некоторое алгебраическое действие (*сложить*, *умножить*, *степень* и т. п.), код этого действия с ранее введенным выражением записывается в одну из рабочих ячеек. Далее происходит формирование второй части фразы аналогично описанному выше с пересылкой информации о функции и аргументе в другие рабочие ячейки. При поступлении слова *вывести* начинается анализ предложения в целом.

Блок-схема алгоритма анализа фраз и выполнения содержащихся в них действий приведена на рис. 60. Основные этапы работы этого алгоритма:

1. Определяется, если ли во фразе алгебраическое действие типа K , что соответствует наличию как $f(x)$, так и $\psi(y)$.

2. При выполнении условия (1) происходит анализ второй половины фразы в соответствии с пп. 3 и 4. Если это условие не выполняется (произнесена фраза типа $f[g(x)]$), осуществляется переход к п. 3.

3. Определяется, не сложная ли функция должна быть вычислена. Если внешняя или внутренняя функции являются числами (например, при выражении $a \sin x$), их переводят из двоично-десятичной в двоичную форму записи.

4. Переводится в двоичную форму записи аргумент. В соответствии с кодами функций вызываются необходимые стандартные программы, по которым происходит вычисление внутренней функции, затем внешней.

5. В случае отсутствия дальнейшего алгебраического действия (условие (1) не выполняется) исходные данные и результаты вычислений выдаются на печать с одновременной очисткой рабочих ячеек памяти и восстановлением модифицированных команд.

6. В случае наличия второй части предложения выполняется алгебраическое действие и выдача исходных данных и результата вычислений на печать.

Рассмотрим работу алгоритма на примере ввода простого устного задания на вычисление выражения $\sqrt{1,357}$ (аналогичного использованному для иллюстрации в предыдущем параграфе, но без логарифмирования подкоренного числа). Фраза состоит из слов: *слушай* вычислить корень один точка три пять семь стоп вывести. Блок проверки начального слова *слушай* (см. рис. 59) подготавливает программу к поступлению всей фразы, смысл которой должен быть определен после произнесения последнего слова

фразы *вывести*. После поступления слова *вычислить* и его опознавания определяется тип фразы.

Порядок использования смысловой информации поясним на примере следующего слова фразы *корень*. Если это слово опознано как одно из слов 3 класса, выдается информация, что код функции принят. В случае, если поступившая реализация *корень* опознана как слово другого класса, но близкое в пространстве словесных признаков к какому-либо слову 3 класса, последнее считается поступившим на вход, а код его определяется по таблицам корректировки кодов, содержащихся в программе опознавания фраз. В случае отсутствия опознанного слова в таблице корректировки кодов 3 класса происходит отказ от опознавания и третье слово фразы, в данном случае *корень*, надо повторить.

Далее программа определяет, не поступает ли слово *скобка* (признак наличия сложной функции). В данной фразе вычисляется простая функция, и четвертое слово фразы должно относиться к аргументу. Вначале идет проверка знака числа (не является ли *минус* первым словом аргумента). При поступлении слова *один* программа определяет, что знак числа положительный. Определение знака порядка устанавливается проверкой, не является ли нулем первая цифра аргумента. Программа начинает формировать двоично-десятичное число аргумента. Одновременно ведется счет цифр до слова *точка*, определяющих порядок, который в нашем примере положительный. Каждая следующая цифра аргумента засылается в предназначенную для нее тетраду мантиссы. При появлении слова *точка* формирование порядка прекращается, образование аргумента продолжается, пока не поступит слово *стоп*.

После произнесения и опознавания слова *стоп* проверяется, не является ли данная функция тригонометрической. Следующее слово, поступившее на вход, — *вывести*, означает, что фраза закончена и необходимо перейти к ее анализу (рис. 60). В данном примере фраза не содержит вычисления сложной функции и алгебраических действий. После этих проверок аргумент печатается и затем переводится из двоично-десятичной системы в двоичную. Далее, в соответствии с кодом функции, программа формирует команду извлечения квадратного корня, после чего выполняет эту команду, переводит результат в двоично-десятичную форму и печатает его. Примерно через 0,5 сек после окончания произнесения последнего слова фразы печатающее устройство ЦВМ выдало исходные данные и результат вычисления в следующем виде:

037	000	000	000	000	— код слова <i>слушай</i>
057	000	000	000	000	— код слова <i>вычислить</i>
035	000	000	000	000	— код слова <i>корень</i>
++	01	135	700	000	— аргумент
++	01	116	490	342	— результат вычислений

Приведем образцы ответов ЦВМ при вычислении более сложных выражений:

$$1) e^{4,7} \cdot \sqrt[7]{7,5} = 301,102732 \quad 2) [\arctg(\sin 45^\circ)]^{\sqrt[7]{\ln 1,7}} = 0,702189549$$

037	000	000	000	000	037	000	000	000	000
057	000	000	000	000	057	000	000	000	000
027	000	000	000	000	060	000	000	000	000
++	01	470	000	000	016	000	000	000	000
013	000	000	000	000	++	00	450	000	000
035	000	000	000	000	052	000	000	000	000
++	01	750	000	000	035	000	000	000	000
++	03	301	102	732	017	000	000	000	000
					++	01	170	000	000
					++	00	702	189	549

Наименования кодов даны в табл. 17.

Результаты эксперимента

Эксперименты по речевому управлению ЦВМ БЭСМ-3М с помощью описанного алгоритма проводились на фразах, содержащих задания на вычисление различных математических выражений. Микрофон был установлен на пульте ЦВМ. Уровень шума в зале составлял около 60 дб. Для получения количественных оценок были составлены 20 фраз-заданий на вычисление математических выражений типа приведенных выше. Общее число слов во фразах составило 235.

В эксперименте приняло участие 20 лиц обоего пола, произносившие все устные задания. Надежность опознавания фраз составила 90,2% (из общего числа 400 фраз, содержащих 4700 реализаций слов, в 39 фразах было 41 неправильно опознанное слово). Надежность опознавания слов во фразах составила 99,15%, что весьма существенно превысило достигнутую ранее надежность — 93%, полученную при опознавании слов без учета смысловой информации. В результате повторного произнесения указанных 39 фраз 33 из них были приняты правильно, и, таким образом, надежность опознавания фраз возросла до 98,5%. Распределение ошибок при задании функций, алгебраических действий и при формировании аргументов приведены в табл. 20, 21, 22. Ошибки, оставшиеся после повторного произнесения ошибочно опознанных фраз, показаны цифрами в скобках.

При использовании лингвистической информации следует иметь в виду, что иногда два (или более) слова одного класса могут быть опознаны как одно из слов, не входящих в этот класс. В этом случае следует переходить к алгоритму различения лишь этих двух (или более) слов. В случае контроля за правильностью опознавания слов во фразах с помощью системы отображения инфор-

Т а б л и ц а 20

Распределение ошибок при задании функции

Произнесено	Опознано										Общее число слов
	синус	косинус	тангенс	котангенс	арксинус	арккосинус	арктангенс	корень	логарифм	экспонента	
синус	80										80
косинус		60									60
тангенс			60								60
котангенс				39			1				40
арксинус					39		1				40
арккосинус						40					40
арктангенс				3 (1)		2	35				40
корень								98	2 (1)		100
логарифм			4 (2)					4	72		80
экспонента										60	60

Т а б л и ц а 21

Распределение ошибок при задании алгебраических действий

Произнесено	Опознано					Общее число слов
	сложить	вычесть	умножить	разделить	степень	
сложить	20					20
вычесть		40				40
умножить			40			40
разделить		1	1	18		20
степень					40	40

мации на электроннолучевой трубке появляется возможность исправлять повторным произнесением ошибочно опознанные слова, не дожидаясь окончания фразы.

Описанная система речевого управления находится длительное время в действии и используется в экспериментальных и демонстрационных целях. Опыт речевого управления машиной посредством устных заданий показал, что таким образом можно удобно и весьма быстро получать ответы на поставленные вопросы и решения различных задач. Эксперимент, описанный в данном параграфе, проводился автором совместно с Высоцким, Рудным и Труниным-Донским [20].

Т а б л и ц а 22

Распределение ошибок при формировании аргумента

Произне- сено	Опознано												Общее число слов	
	ноль	один	два	три	четыре	пять	шесть	семь	восемь	девять	точка	минус		стоп
ноль	136		4 (1)											140
один		137							3					140
два	7 (2)		113											120
три				140										140
четыре					140									140
пять						159					1			160
шесть							118		1	1				120
семь								140						140
восемь	1								119					120
девять						1				99				100
точка											320			320
минус						1						19		20
стоп					2								558	560

На основе принципов, изложенных выше, разработана система речевого ввода основных параметров газопроводов в ЦВМ для оперативного расчета эксплуатационного режима (Рудный, Трунин-Донской в сотрудничестве с Максимовым и Сакаевым). Словарь состоит из 33-х слов. Осуществление визуального контроля на пульте оператора позволяет повторять каждое неправильно опознанное слово после произнесения слова *стереть*. Вводятся 8 параметров (диаметр трубы газопровода, температура воздуха, давление газа на входе и выходе и др.). Фразы-задания на вычисление режима газопроводов состоят из 50—55 слов. Порядок ввода отдельных параметров произвольный.

ЗАКЛЮЧЕНИЕ

Дальнейшие исследования по проблеме автоматического опознавания речи должны проводиться на всех уровнях, начиная от увеличения числа и вариантов опознаваемых сегментов, фоном и слов и кончая введением в более значительной степени лингвистической информации при опознавании фраз. На пути инженерного решения этих задач имеются большие технические трудности. От того, насколько удастся преодолевать эти трудности, зависят возможности опознающих машин. В работе над данной проблемой (как и при решении многих других сложных задач) приходится рассматривать отдельные частные случаи, *сознательно* допускать упрощения структуры речевого процесса, моделируемого в той или иной степени в опознающих машинах. Эти упрощения еще длительное время неизбежны, когда речь идет о моделировании процессов, происходящих в наиболее высокоорганизованном организме — человеке.

Развитие исследований по проблеме автоматического опознавания речи зависит в значительной мере от практического использования достигнутых к настоящему времени результатов. Наряду с системами, опознающими десятки и сотни слов, могут найти применение устройства со словарём в несколько слов. По мнению Голда [157], наиболее обещающей ближайшей областью применения опознавания слов является речевая связь с вычислительной машиной. «Многие специалисты по вычислительной технике считают, что телетайпная связь с ней более надежна и требует меньшую память. Подобный аргумент был когда-то очень убедителен, но современные условия, когда человек и машина должны быстро реагировать, сделали его менее убедительным», — пишет Голд далее. По данным [114], в проекте машины с разделением времени Genie, разработанном в Калифорнийском университете (Беркли), для исследования вопросов взаимодействия в системе человек — машина наряду с несколькими телетайпными входами предусмотрены два речевых входа.

Пока еще нет сведений о практическом применении разработанных устройств управления голосом. Между тем, только опыт практического использования речевого управления может ответить на ряд вопросов, таких, как целесообразность установления речевой связи в различных условиях, установление разумного compro-

мисса между требованием машины говорить более четко с некоторым приспособлением к возможностям ее «восприятия» и привычкой человека к разговорному стилю речи, рассчитанному на понимание «с полуслова».

Рассматривая последний вопрос, акад. Харкевич указывал, что нельзя требовать, чтобы при любом произнесении машина воспринимала все команды безошибочно. Чем больше произвол, допускаемый при произнесении тех или иных слов, тем меньше вероятность их опознавания данной машиной или тем труднее построить машину, осуществляющую опознавание с заданной надежностью. Но, с другой стороны, перекладывая трудности с машины на человека, можно не облегчить, а усложнить труд оператора. В расширенной постановке проблема организации взаимодействия человека и машины состоит в обеспечении заданной высокой надежности при минимальных затратах на аппаратуру и максимально комфортабельных условиях работы оператора [94].

В давние времена люди мечтали о таких, казалось бы, несбыточных вещах, как ковер-самолет; зеркало, в котором видно происходящее в далеких странах; дверь, которую *любой* может открыть словами: «Сезам, отворись». Прошли годы и столетия, многое из этих мечтаний стало привычным и будничным, но вещи или машины, которые слушались бы голоса человека, пока в обиход не вошли. Надо полагать, что в ближайшие годы речевая связь с машиной начнет получать распространение не только в произведениях писателей-фантастов, но и в реальной жизни.

ПРИЛОЖЕНИЕ

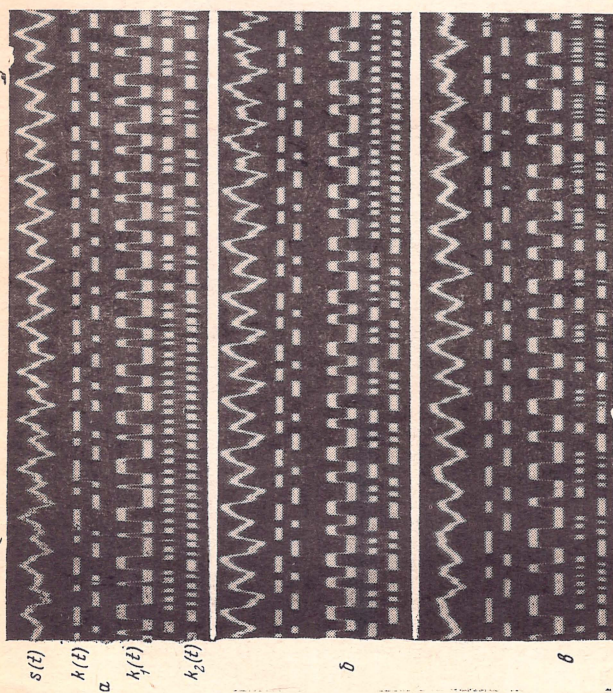
На рис. 61—64 приведены осциллограммы речевого сигнала до и после клиппирования — $s(t)$ и соответственно $k(t)$, сигналов на выходах цепей, выделяющих приближенные значения первой и второй формант — $k_1(t)$ и $k_2(t)$, для отрезков отдельных фонем и звукосо сочетаний. Число импульсов сигналов $k_1(t)$ и $k_2(t)$ пропорционально усредненным значениям F_1 и F_2 . Масштаб времени осциллограмм 1 мм—0,75 мсек.

На рис. 61 приведены осциллограммы для начальной части y (включающей переходный сегмент и часть квазистационарного) из слова *шурф*, конечной части y из слова *шутят* и начального переходного сегмента y из слова *выжил*. Осциллограммы имеют существенные различия лишь по кривой $k_2(t)$. Переходные сегменты y обладают повышенным значением F_2 . Переходный сегмент y типичен для сочетаний y с губными согласными. F_2 в нем нарастает от ~ 700 до ~ 1700 гц.

На рис. 62 приведены осциллограммы для начальных частей o в словах *кочка*, *точка* и переход от c к $и$ в слове *синус*. На кривых $k_2(t)$ и $s(t)$ хорошо видно повышенное значение F_2 в переходном сегменте o слова *точка*. Заметно запаздывание F_2 по сравнению с F_1 в сочетании *си*.

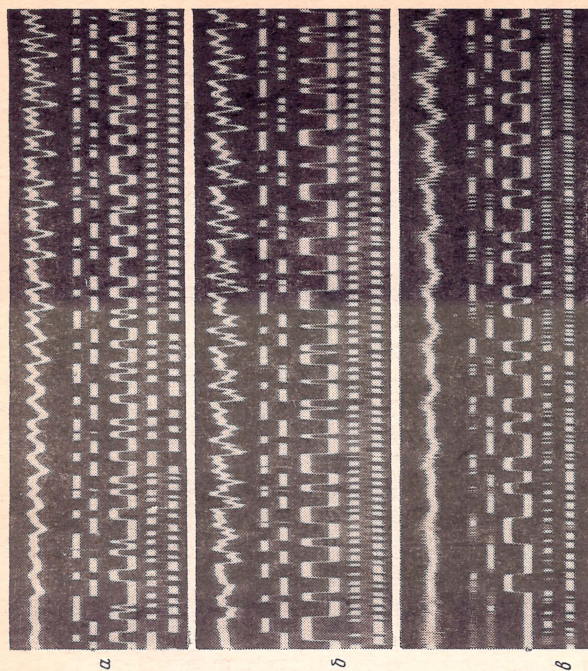
На рис. 63 приведены осциллограммы для начальных частей «мягких» гласных 'э' из слова *семь*, 'а' из слова *пять* и jy из слога *юм*. Параметры 'э' изменяются мало. Начальная часть 'а' характеризуется постепенным увеличением F_1 , сопровождаемым уменьшением F_2 .

На рис. 64 приведены осциллограммы $s(t)$ и $k(t)$ для c из слова *синус*, перехода от $ш$ к y из слова *шурф*, x из слова *хата*, k из слова *бок* и перехода от 'э' к m' из слова *семь*.



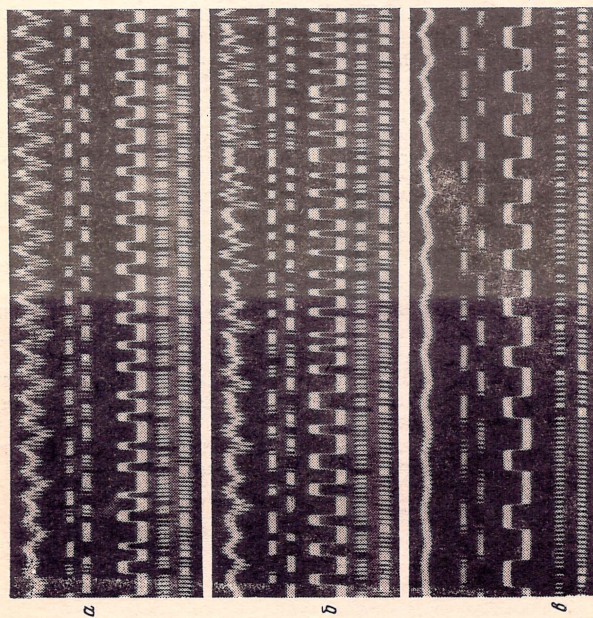
Р и с. 61

Осциллограммы отрезков речевого сигнала до и после клиппирования для гласных y (кривые a и b') и y (кривые $в$)



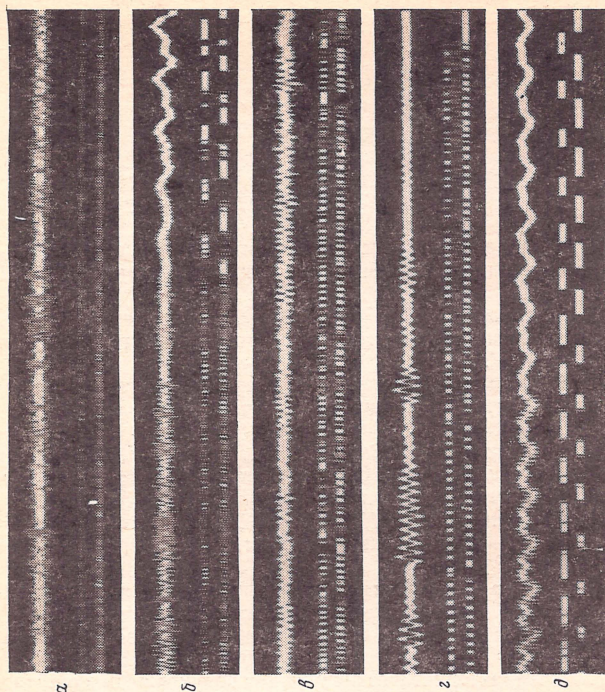
Р и с. 62

Осциллограммы отрезков речевого сигнала до и после клиппирования для гласных o (кривые a и $б'$) и сочетания ci (кривые $в$)



Р и с. 63

Осциллограммы отрезков речевого сигнала до и после клипирования для гласных 'э' (кривые *а*), 'а' (кривые *б*) и сочетания *ју* (кривые *в*)



Р и с. 64

Осциллограммы отрезков речевого сигнала для *с'* (кривые *а*) сочетания *шу* (кривые *б*), согласных *х* и *к* (кривые *в* и *г*) и сочетания *ешк* (кривые *д*)

ЛИТЕРАТУРА

1. Акинфиев Н. Н. К вопросу построения теории речевых сообщений. Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1957.
2. Барабаш Ю. Л., Варский Б. В., Зиновьев В. Т., Кириченко В. С., Сапегин В. Ф. Вопросы статистической теории распознавания, гл. 14. М., изд-во «Советское радио», 1967.
3. Бондарко Л. В. О характере изменения формантного состава русских гласных под влиянием мягкости соседних согласных.— Уч. зап. ЛГУ, 1960, № 237, вып. 40.
4. Бондарко Л. В., Вербицкая Л. А., Зиндер Л. Р., Павлова Л. П. Различаемые звуковые единицы русской речи.— Сб. «Механизмы речеобразования и восприятия сложных звуков». М.—Л., изд-во «Наука», 1966, 165—179.
5. Бондарко Л. В., Загоруйко Н. Г., Кожеевников В. А., Молчанов А. П., Чистович Л. А. Модель восприятия речи человеком. Новосибирск, изд-во «Наука», 1968.
6. Бондарко Л. В., Зиндер Л. Р. О некоторых дифференциальных признаках русских согласных фонем.— Вопр. языкознания, 1966, № 1, 10—14.
7. Быков Ю. С. Теория разборчивости речи и повышение эффективности радиотелефонной связи. М.—Л., Госэнергоиздат, 1959.
8. Варшавский Л. А. Характеристические признаки звуков речи и перспективы предельной частотной компрессии.— Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1957.
9. Варшавский Л. А. Значение статистических исследований речи для техники.— Сб. «Вопросы статистики речи». ЛГУ, 1958, 5—11.
10. Варшавский Л. А., Литвак И. М. Исследование формантного состава и некоторых других физических характеристик звуков русской речи.— Сб. «Проблемы физиологической акустики», вып. 3. М.—Л., Изд-во АН СССР, 1955, 5—17.
11. Величко В. М., Загоруйко Н. Г. Автоматическое распознавание ограниченного набора устных команд.— Сб. «Вычислительные системы», вып. 36. Новосибирск, 1969, 101—110.
12. Винцюк Т. К. Алгоритм оптимального членения речевого сигнала на части (сегменты).— Тр. IV Всесоюзн. школы-семинара «Автом. распознавание слуховых образов». Киев, 1969, 135—142.
13. Винцюк Т. К. Распознавание слов устной речи методами динамического программирования.— Кибернетика, 1968, № 1, 81—88.
14. Волошин Г. Я. Об использовании языковой избыточности для повышения надежности автоматического распознавания речевых сигналов.— Сб. «Вычислительные системы», вып. 28. Новосибирск, изд-во «Наука», 1967, 21—48.
15. Волошин Г. Я., Бахмутова И. В., Прокопенко А. А. О сетевом алгоритме распознавания фонем по последовательности сегментов.— Сб. «Вычислительные системы», вып. 37. Новосибирск, 1969, 44—47.
16. Высоцкий В. Д., Тетельбаум С. И. Способ повышения артикуляции в канале связи с помехами — Изв. электропром. слабых токов, 1938, № 1.
17. Высоцкий Г. Я., Кулагин М. В., Рудный Б. Н., Трунин-Донской В. Н. Автоматическое опознавание некоторых шумных звуков в произвольном контексте.— Докл. на юбилейной конференции. ВЦ АН СССР, 1970.

18. *Высоцкий Г. Я., Рудный Б. Н., Трунин-Донской В. Н., Цемель Г. И.* Алгоритм опознавания 40 слов на ЦВМ БЭСМ-3М.— Сб. «Работы по технической кибернетике», вып. 2. М., ВЦ АН СССР, 1968, 3—33.
19. *Высоцкий Г. Я., Рудный Б. Н., Трунин-Донской В. Н., Цемель Г. И.* Автоматическое опознавание нескольких десятков слов и фраз, произнесенных произвольным диктором.— Тр. IV Всесоюз. школы-семинара «Автом. распознавание слуховых образов». Киев, 1969, 199—201.
20. *Высоцкий Г. Я., Рудный Б. Н., Трунин-Донской В. Н., Цемель Г. И.* Опыт речевого управления вычислительной машиной.— Изв. АН СССР. Техн. кибернетика, 1970, № 2.
21. *Высоцкий Г. Я., Рудный Б. Н., Трунин-Донской В. Н., Цемель Г. И.* Реализация иерархического автомата для распознавания смысла речевого сообщения.— Тр. Акустического ин-та, 1970, вып. 12.
22. *Высоцкий Г. Я., Трунин-Донской В. Н.* Речевой ввод набора команд в вычислительную машину БЭСМ-3М. Работы по технической кибернетике, вып. 1. М., ВЦ АН СССР, 1967, 52—59.
23. *Высоцкий Г. Я., Трунин-Донской В. Н., Цемель Г. И.* Опознавание 40 слов с помощью цифровой вычислительной машины.— Докл. на VI Всесоюз. акуст. конф. М., февраль 1968.
24. *Галунов В. И., Чистович Л. А.* О связи моторной теории с общей проблемой распознавания речи.— Акуст. ж., 1965, т. 11, вып. 4, 417—426.
25. *Гармаш В. А., Кириллов Н. Е.* Экспериментальное исследование статистических характеристик ограниченной речи. Докл. на Всесоюз. совещании секции речи Комиссии по акустике АН СССР, апрель 1958.
26. *Голубцов С. В.* О применении ЭЦВМ для исследования методов автоматического распознавания речевых элементов.— Тр. IV Всесоюз. школы-семинара «Автом. распознавание слуховых образов». Киев, 1969, 73—86.
27. *Григорян А. А.* О характеристиках каналов для выделения группы гласных звуков и частоты второй форманты.— Тр. Акустического института, 1970, вып. 12.
28. *Григорян А. А.* Исследование динамики формантных частей гласных звуков (в печати).
29. *Григорян А. А., Крюков С. Н., Цемель Г. И.* Опознавание набора слов с использованием признаков переходных сегментов гласных (в печати).
30. *Григорян А. А., Цемель Г. И.* К опознаванию согласных звуков и неоднородных гласных по переходным сегментам гласных.— Проблемы передачи информации, 1969, т. 5, № 4, 38—49.
31. *Гумецкий Р. Я., Дергач М. Ф., Мишин Л. Н., Оверченко М. М., Чабан М. Е.* Распознавание слов из ограниченного алфавита фонем в модели с иерархической структурой.— Тр. Акустического ин-та, 1970, вып. 12.
32. *Дергач М. Ф.* Статистика восприятия глухих взрывных и щелевых в зависимости от их длительности.— Сб. «Вопросы статистики речи». ЛГУ, 1958, 40—44.
33. *Дергач М. Ф.* Статистика временного распределения участков характерного звучания гласных звуков русского языка.— Сб. «Вопросы статистики речи». ЛГУ, 1958, 36—39.
34. *Дукельский Н. И.* Синтез русской речи из отдельных звуков и ее фоноидальный состав.— Докл. на Всесоюз. совещании секции речи Комиссии по акустике АН СССР, апрель 1958.
35. *Дукельский Н. И.* Принципы сегментации речевого потока. М.—Л., Изд-во АН СССР, 1962.
36. *Елкина В. Н., Юдина Л. С.* Статистика слогов русской речи.— Сб. «Вычислительные системы», вып. 10. Новосибирск, 1964, 55—91.
37. *Елкина В. Н., Юдина Л. С., Хайретдинова А. Г.* Статистика двух- и трехфонемных сочетаний русской речи.— Сб. «Вычислительные системы», вып. 37. Новосибирск, 1969, 48—72.
38. *Загоруйко Н. Г.* Об обмене устной информацией между человеком и вычислительными системами.— Сб. «Вычислительные системы», вып. 10. Новосибирск, 1964, 3—12.

39. *Загоруйко Н. Г.* Алгоритм распознавания фонем по последовательности сегментных решений.— Тр. IV Всесоюзн. школы-семинара «Автом. распознавание слуховых образов». Киев, 1969, 143—146.
40. *Загоруйко Н. Г., Величко В. М., Волошин Г. Я.* и др. Эксперименты по автоматическому распознаванию речевых сигналов.— Тр. IV Всесоюзн. школы-семинара «Автом. распознавание слуховых образов». Киев, 1969, 147—151.
41. *Загоруйко Н. Г., Волошин Г. Я., Ёлкина В. Н.* Автоматическое опознавание звуковых образов (обзор литературы). Сб. «Вычислительные системы», вып. 14. Новосибирск, 1964, 3—30.
42. *Зайцев В. Г., Тимофеев Б. Б.* Распознавание клиппированной речи с помощью вычислительной машины.— Автоматика и приборостроение. Информ. научно-техн. сб., 1965, № 2 (22), 19—22.
43. *Зиндер Л. Р.* Общая фонетика. Изд. ЛГУ, 1960.
44. *Златоустова Л. В.* Фонетические единицы речи. Изд-во МГУ (в печати).
45. *Какауридзе А. Г.* Некоторые вопросы кодирования гласных звуков речи.— Тр. Ин-та электроники, автом. и телемех. АН Груз.ССР, 1960, т. 1, 41—63.
46. *Какауридзе А. Г.* Некоторые вопросы автоматического распознавания речевых сигналов.— Тр. Ин-та электроники, автом. и телемех. АН Груз.ССР, 1963, т. 4.
47. *Какауридзе А. Г.* Экспериментальное устройство для автоматического различения ограниченного набора речевых команд.— Сб. «Элементы вычислительной техники и машинный перевод». Тбилиси. Изд-во «Мецниереба», 1964, 143—163.
48. *Какауридзе А. Г., Доценко Г. С.* Экспериментальное исследование возможности автоматического распознавания речевых команд. Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1959.
49. *Какауридзе А. Г., Рамшвили Г. С.* О роли звуков речи в узнавании голосов.— Сб. «Элементы вычислительной техники и машинный перевод». Тбилиси. Изд-во «Мецниереба», 1964, 165—169.
50. *Кандинов А. В., Цемель Г. И.* О возможности сужения спектра телефонного канала передач мгновенной частоты речи или применением делителей частоты.— Электросвязь, 1958, № 8, 3—8.
51. *Кириченко В. С.* Способ регистрации измерений текущего спектра речи. Авт. свид. № 139004.— Бюлл. изобр., 1961, № 12.
52. *Книппер А. В., Петров О. А.* Опыт опознания небольшого набора слов.— Сб. «Опознание образов». М., изд-во «Наука», 1968, 155—165.
53. *Книппер А. В., Турбович И. Т., Юрков Е. Ф.* Выбор метода и объекта опознания речевых образов на основе специфики речи.— Сб. «Опознание образов». М., изд-во «Наука», 1968, 86—97.
54. *Книппер А. В., Турбович И. Т., Юрков Е. Ф.* Принципы построения аналогового опознающего устройства, реализующего детерминированный метод опознания с учетом статистических свойств выборки.— Сб. «Опознание образов». М., изд-во «Наука», 1968, 98—124.
55. *Кринов С. Н.* Устройство для выделения двух градаций основного тона.— Сб. «Работы по технической кибернетике», вып. 3. Изд-во ВЦ АН СССР, 1970.
56. *Кузнецова А. М.* Изменения гласных под влиянием соседних мягких согласных. М., изд-во «Наука», 1965.
57. *Кюнnap Э.* Устные команды в системах управления.— Изв. АН ЭССР, сер. физ.-матем. и техн. наук, т. 15, № 3, 377—399.
58. *Лев А. Ю.* О спектрах модуля и фазы аналитического сигнала с ограниченным спектром. Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1959.
59. *Линдгрен Н.* Машинное распознавание человеческого языка.— Зарубежная радиоэлектроника, 1966, № 3, 55—89; № 4, 63—89; № 5, 56—76.
60. *Лисенко Д. М.* О членении человеком непрерывного потока речи на слова. Сб. «Механизмы речеобразования и восприятия сложных звуков». М.—Л., изд-во «Наука», 1966, 180—192.

61. Лийв Г., Ээк А. О проблемах экспериментального изучения динамики речеобразования: комплексная методика синхронизированного кинофлуорографирования и спектрографирования речи.— Изв. АН ЭССР, Биология, 1968, т. 17, № 1, 78—102.
62. Лобанов Б. М. Об автоматическом выделении шумных звуков по сигналам клипированной речи.— Электросвязь, 1968, № 11, 59—64.
63. Люблинская В. В. Распознавание артикуляторных признаков смычных согласных по переходу от гласного к согласному.— Акуст. ж., 1966, т. 12, вып. 2, 213—221.
64. Мартынов В. С. Новая система выделения основного тона. Докл. на Всесоюз. совещании секции речи Комиссии по акустике АН СССР, апрель 1958.
65. Матусевич М. И. Введение в общую фонетику. М., Учпедгиз, 1959.
66. Мясников Л. Л. Объективное распознавание звуков речи.— Журн. техн. физ., 1943, т. 13, № 3, 109—115.
67. Мясников Л. Л. Звуки речи и их объективное распознавание.— Вестник ЛГУ, 1946, № 3.
68. Мясников Л. Л. Физические исследования звуков русской речи.— Изв. АН СССР, серия физическая, 1949, т. 13, № 6, 723—727.
69. Мясникова Е. Н. Объективное распознавание звуков речи. Л., изд-во «Энергия», 1967.
70. Осадчий Ю. Н. Оценка возможности распознавания ограниченного набора команд с использованием субфонемных последовательностей.— Тр. Акустического ин-та, 1970, вып. 12.
71. Петров О. А. Статистическая обработка первичных признаков речевых сигналов с использованием ЭВМ.— Сб. «Опознавание образов. Теория передачи информации». М., изд-во «Наука», 1965, 30—36.
72. Пирогов А. А. К вопросу о фонетическом кодировании речи.— Электросвязь, 1967, № 5, 24—31.
73. Покровский Н. Б. Расчет и измерение разборчивости речи. Связьиздат, 1962.
74. Рамишвили Г. С. Об автоматическом узнавании голосов.— Изв. АН СССР, Техн. кибернетика, 1966, № 5, 87—92.
75. Рамишвили Г. С., Какауридзе А. Г. Об узнаваемости голосов в условиях ограничения речевого сигнала по спектру и амплитуде.— Сообщения АН Груз. ССР, 1965, т. 40, 39—45.
76. Ростовцев Ю. Г. О законах распределения нулевых и экстремальных точек сигналов русской речи при сильном ограничении по амплитуде.— Радиотехника, 1958, т. 13, № 4, 63—67.
77. Ростовцев Ю. Г. О возможности применения в системах связи предельного амплитудного ограничения речевых сигналов.— Электросвязь, 1958, № 6, 49—52.
78. Сапожков М. А. Речевой сигнал в кибернетике и связи. М., Связьиздат, 1963.
79. Сапожков М. А. О методах «компрессии» спектров речи.— Электросвязь, 1958, № 8, 36—47.
80. Сорокин В. Н., Файн В. С. Непрерывно-групповое распознавание слов: алгоритм и экспериментальные результаты. Докл. на VI Всесоюз. акуст. конференции. М., февр. 1968.
81. Трунин-Донской В. Н. Об использовании лингвистической информации при автоматическом опознавании фраз.— Сб. «Работы по технической кибернетике», вып. 3. М., Изд-во ВЦ АН СССР, 1970.
82. Трунин-Донской В. Н., Фирер А. С., Цемель Г. И. Опознавание набора слов с помощью цифровой вычислительной машины.— Сб. «Работы по технической кибернетике», вып. 1. М., Изд-во ВЦ АН СССР, 1967, 37—51.
83. Трунин-Донской В. Н., Цемель Г. И. Опознавание гласных звуков по клипированному речевому сигналу.— Проблемы передачи информации, 1968, т. 4, № 2, 62—70.
84. Трунин-Донской В. Н., Цемель Г. И. Опознающее устройство для речевого ввода данных в вычислительную машину. Авт. свид. № 251270.— Бюлл. изобр., 1969, № 27.

85. Турбович И. Т. Об умножении и делении частоты немодулированных и модулированных колебаний.— Радиотехника, 1956, № 8.
86. Турбович И. Т., Петров О. А. Об одном методе полного описания одномерных образов совокупностью простых функций (применительно к речевым сигналам).— Сб. «Опознавание образов. Теория передачи информации». М., изд-во «Наука», 1965, 25—29.
87. Уггер В. Г. Некоторые свойства косинус — фи составляющей речевых сигналов. Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1958.
88. Фант Г. Акустическая теория речеобразования. М., изд-во «Наука», 1964.
89. Файн В. С. Опознавание изображений (основы непрерывно-групповой теории и ее приложения), гл. 8. М., изд-во «Наука», 1970.
90. Файн В. С., Сорокин В. Н. К вопросу о распознавании слогов на основе «видимой речи».— Проблемы передачи информации, 1967, т. 3, вып. 3, 64—72.
91. Фланнаган Дж. Анализ, синтез и восприятие речи. М., изд-во «Связь», 1968.
92. Харкевич А. А. Опознавание образов.— Радиотехника, 1959, т. 14, № 5, 12—22.
93. Харкевич А. А. О принципах построения читающих машин.— Радиотехника, 1960, т. 15, № 2.
94. Харкевич А. А. Некоторые методические вопросы в проблеме опознавания.— Проблемы передачи информации, 1965, т. 1, вып. 3, 3—11.
95. Цемель Г. И. О соотношении спектров речевого сигнала и его мгновенной частоты, необходимых для передачи речи. Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1958.
96. Цемель Г. И. К определению инвариантных признаков смычных (взрывных) звуков по сигналам клиппированной речи.— Изв. АН СССР, ОТН, Энергетика и автоматика, 1959, № 4, 214—215.
97. Цемель Г. И. Об использовании сигналов клиппированной речи для объективного распознавания звуков речи. Докл. на Всесоюзн. совещании секции речи Комиссии по акустике АН СССР, апрель 1959.
98. Цемель Г. И. Автоматическое различение ограниченного набора слов.— Изв. АН СССР, ОТН, Энергетика и автоматика, 1960, № 5, 179—182.
99. Цемель Г. И. Объективное распознавание ограниченного набора звуков и слов.— Сб. «Проблемы передачи информации», вып. 10. М., Изд-во АН СССР, 1961, 57—62.
100. Цемель Г. И. Автоматическое (объективное) распознавание звуков речи (обзор).— Зарубежная радиоэлектроника, 1961, № 4, 52—73.
101. Цемель Г. И. Повышение надежности объективного распознавания звуков речи введением переспроса.— Сб. «Системы распределения информации. Опознавание образов». М., Изд-во АН СССР, 1963, 77—79.
102. Цемель Г. И. Опознавание набора слов по характерным признакам речевого сигнала.— Сб. «Теория передачи информации». М., изд-во «Наука», 1964, 57—68.
103. Цемель Г. И. Опознавание глухих щелевых звуков по клиппированному речевому сигналу.— Проблемы передачи информации, 1965, т. 1, № 4, 33—40.
104. Цемель Г. И. О признаках речевого сигнала на уровне сегментов.— Тр. Акустического ин-та, 1970, вып. 12.
105. Чистович Л. А. Применение статистических методов к определению фонетической принадлежности индивидуального гласного звука.— Сб. «Вопросы статистики речи». ЛГУ, 1958, 26—35.
106. Чистович Л. А. Текущее распознавание речи человеком.— Сб. «Машинный перевод и прикладная лингвистика». М., I — вып. 6, 1961, 37—79; II — вып. 7, 1962, 3—44.
107. Чистович Л. А. Изменение основной частоты голоса как различительный признак согласных.— Акуст. ж., 1968, т. 14, № 3, 449—456.
108. Чистович Л. А., Клаас Ю. А., Кузьмин Ю. И. Текущее распознавание звуков речи (Опыты с быстрым повторением и буквенной записью согласных).— Вopr. психологии, 1962, № 6, 26—29.

109. Чистович Л. А., Кожевников В. А., Алякринский В. В. и др. Речь. Артикуляция и восприятие. М.—Л., изд-во «Наука», 1965.
110. Чистович Л. А., Кожевников В. А. Восприятие речи.— Сб. «Вопросы теории и методов исследования восприятия речевых сигналов». Информ. материалы, вып. 22. Изд. Объединенного научн. совета «Физиология человека и животных», 1969, 4—149.
111. Шакин В. В. Узнавание голосов на основе аппроксимации квантованной видимой речи.— Сб. «Опознавание образов». М., изд-во «Наука», 1968, 207—215.
112. Щерба Л. В. Избранные работы по русскому языку. М., Учпедгиз, 1957.
113. Юрков Е. Ф. Полуалгоритмический метод нахождения вторичных признаков, соответствующих сегментам речи.— Сб. «Опознавание образов». М., изд-во «Наука», 1968, 125—131.
114. Системы с разделением времени (под ред. У. Карплюса). М., изд-во «Мир», 1969.
115. Трубецкой Н. С. Основы фонологии. М., ИЛ, 1960.
116. Ahmed R. Vowel recognition in clipped speech.— *Nature*, 1958, v. 181, N 4603, 218.
117. Anke D., Hoeschele P. Einfache Erkennungsgeräte für die gesprochenen Zahlen NULL bis NEUN.— *Kybernetik*, 1968, Bd.4, N 6, 228—234.
118. Bastian J., Delattre P. C., Liberman A. M. Silent interval as a cue for the distinction between stops and semivowels in media position.— *JASA*, 1959, v. 31, N 11, 1568.
119. Berkeley E. C. Translating spoken english into written words.— *Computers and Automat.*, 1956, v. 5, N 3.
120. Bezdel W., Chandler H. J. Results of an analysis and recognition of vowels by computer using zero-crossing data.— *Proc. IEE*, 1965, v. 112, N 11, 2060—2066.
121. Bezdel W., Hawken R. E. Human factors in the transmission of data by speech using automatic speech recognition.— *PTT — bedrijf*, 1967, v. 15, N 1—2, 207—211.
122. Black J. W., Hixson W. Number of axis crossing and the intelligibility of speech.— *JASA*, 1959, v. 31, N 10, 1384—1385.
123. Bliss J. C. Will sound control machines.— *Automat. Control*, 1960, v. 12, N 1.
124. Bogert B. P. On the band width of vowel formants.— *JASA*, 1953, v. 25, N 4, 791—792.
125. Chang S.-H., Dolansky L. O., Howard C. R. A formant Tracker for speech sounds.— *JASA*, 1956, v. 28, N 1, 161.
126. Chang S.-H., Pihl G. E., Essigmann M. W. Representation of speech sounds and some of their statistical properties.— *Proc. IRE*, 1951, v. 39, N 2, 147—153.
127. Cherry E. C., Phillips V. I. Some possible uses of single sideband signals in formant-tracking systems.— *JASA*, 1961, v. 33, N 8, 1067—1077.
128. Christov P. An experimental joint distribution indicator; possible application for training the deaf.— *Acustica*, 1969, v. 21, N 1, 36—40.
129. Christov P. Possibilities of fricative and stop consonants recognition by waveform parameters.— *IV Budapesti akuszt. konf.*, 1967.
130. Comer D. J. The use of waveform asymmetry to identify voiced sounds.— *IEEE Trans. Audio and Electroacoust.*, 1968, v. AU-16, N 4, 500—506.
131. Cooper F. S., Delattre P. C., Liberman A. M., Borst I., Gerstman L. Some experiments on the perception of synthetic speech sounds.— *JASA*, 1952, v. 24, N 6, 597—606.
132. Davis K. H., Biddulph R., Balashek S. Automatic recognition of spoken digits.— *JASA*, 1952, v. 24, N 6, 637—642.
133. Delattre P. C., Liberman A. M., Cooper F. S. Acoustic loci and transitional cues for consonants.— *JASA*, 1955, v. 27, N 4, 769—773.
134. Delattre P. C., Liberman A. M., Cooper F. S. Formant transition and loci as

- acoustic correlates of place of articulation in american fricatives.— *Studia linguistica*, 1964, v. 16, N 2, 104—121.
135. *Denes P.* The design and operation on the mechanical speech recognizer.— *J. Brit. IRE*, 1959, v. 19, N 4, 219—229.
 136. *Denes P., Mathews M. V.* Spoken digit recognition using Time-Frequency pattern matching.— *JASA*, 1960, v. 32, N 11, 1450—1455.
 137. *Dersch W. C.* Shoebox — a voice responsive machine.— *Datamation*, 1962, v 8, N 6.
 138. *Doshita S.* Studies on the analysis and recognition of japanese speech sounds. Thesis. Kyoto Univ., September 1965.
 139. *Dreyfus-Graf J.* Sonograph and sound mechanics.— *JASA*, 1950, v 22, 731—739.
 140. *Dreyfus-Graf J.* Phonétographe et subformants.— *Techn. Mitt. PTT*, 1957, N 2, 41—59.
 141. *Dreyfus Graf J.* Phonétographe: present et futur.— *Bull. techn. PTT*, 1961, N 5, 160—170.
 142. *Dudley H. W.* Phonetic pattern recognition for narrow-band transmission.— *JASA*, 1958, v. 30, N 8, 733—739.
 143. *Dudley H., Balashek S.* Automatic recognition of phonetic pattern in speech.— *JASA*, 1958, v. 30, N 8, 721—732.
 144. *Dunn H. K.* Methods of measuring vowel formant bandwidths.— *JASA*, 1961, v. 33, N. 12, 1737—1746.
 145. *Endres W.* Drittes Symposium über informationstheorie.— *NTZ*, 1956, Bd. 9, N 6, 261—265.
 146. *Enkel F.* Über den Informationsgehalt der Geräuschlaute.— *NTZ*, 1956, N 11, 493—498.
 147. *Ewing G. D., Taylor J. F.* Computer recognition of speech using zero-crossing information.— *IEEE Trans. Audio and Electroacoust*, 1969, v. AU-17, N 1, 37—40.
 148. *Fant G., Lindblom B.* Studies of minimal speech sound units.— *Quart. Progr. and Status Rept Speech Transmiss. Lab.*, 1961, N 2, 1—11.
 149. *Fatenchand R.* Machine recognition of spoken words. Advanced in computers, v. 1. N. Y.— London, Acad. Press, 1960, 193—229.
 150. *Favella L. F., Reineri M. T., Righini G. U.* The probabilistic classification of time sequences and its application to the recognition of human voice.— *Nuovo cimento*, 1965, v. 36, N 3, 1023—1034.
 151. *Flanagan J. L.* Automatic extraction of formant frequencies from continuous speech.— *JASA*, 1956, v. 28, N 1, 110—118.
 152. *Forgie J. W., Forgie C. D.* Results obtained from a vowel recognition computer program.— *JASA*, 1959, v. 31, N 11, 1480—1489.
 153. *Forgie J. W.* A computer program for recognizing the English fricative consonants *f* and *θ*. 4-th. Internat. Congr. Acoust. Copenhagen, 1962, Congr. Rept, v. 1, s. a., NG 11.
 154. *Foulkes J. D.* Computer identification of vowel types.— *JASA*, 1961 v. 33, N 1, 7—11.
 155. *Fry D. B.* Theoretical aspects of mechanical speech recognition.— *J. Brit. IRE*, 1959, v. 19, N 4, 211—218.
 156. *Gilli L., Meo A. R.* Sequential system for recognizing spoken digits in real time.— *Acustica*, 1967—1968, v. 19, N 1.
 157. *Gold B.* Word-recognition computer program.— *Mass. Inst. Technol. Techn. Rept.* 452, 1966.
 158. *Goodwin P., Bezdel W.* Apparatus for the recognition of speech.— *Brit. Pat.* N 1055371, 18.1.1967.
 159. *Guelke P. W., Smith E. D.* Distribution of information in stop consonants.— *Proc. IEE*, 1963, v. 110, N 2, 680—688.
 160. *Gupta J. P., Agrawal S. S., Ahmed R.* Perception of (Hindi) consonants in clipped speech.— *JASA*, 1969, v. 45, N 3, 770—773.
 161. *Halle M.* The sound pattern of Russian s'Gravenhage, Mouton and Co., 1959.

162. *Halle M., Hughes G. W., Radley P. A.* Acoustic properties of stop consonants.— *JASA*, 1957, v. 29, N 1, 107—111.
163. *Halle M., Stevens K.* Speech recognition: A model and a program for research.— *IRE Trans. Inform. Theory*, 1962, v. IT-8, N 2, 155—159.
164. *Hargreaves W. A., Starkweather J. A.* Recognition of speaker identity.— *Language and Speech*, 1963, v. 6, p. 2.
165. *Harris C. M.* A study of the building blocks in speech.— *JASA*, 1953, v. 25, N 5, 962—969.
166. *Harris K., Hoffman H., Liberman A., Delattre P., Cooper F.* Effect of third-formant transitions on the perception of the voiced stop consonants.— *JASA*, 1958, v. 30, N 2, 122—126.
167. *Hecker M. H. L.* Studies of nasal consonants with an articulatory speech synthesizer.— *JASA*, 1962, v. 34, N 2, 179—188.
168. *Heinz J. M., Stevens K. N.* On the properties of voiceless fricative consonants.— *JASA*, 1961, v. 33, N 5, 589—596.
169. *Hemdal J. F., Hughes G. W.* A feature based computer recognition program for the modeling of vowel perception. Models for the perception of speech and visual form.— *Proc. Sympos. November 1964*, MIT Press. Cambridge, Mass., 1967, 440—453.
170. *Hoffman H. S.* Study of some cues in the perception of the voiced stop consonants.— *JASA*, 1958, v. 30, N 11, 1035—1041.
171. *Hughes G. W., Halle M.* Spectral properties of fricative consonants.— *JASA*, 1956, v. 28, 303—310.
172. *Hughes G. W., Halle M.* On the recognition of speech by machine.— *Proc. Internat. Conf. Inform. Processing. Paris., UNESCO*, June 1959.
173. *Inomata S.* Program for active segmentation and reduction of phonetic parameters. 4-th Internat. Congr. Acoust. Copenhagen, 1962, Congr. Rept, v. 1, s. a. NG 15.
174. *Inomata S.* Speech recognition and generation by a digital computers.— *Res. Electrotechn. Lab.*, 1963, N 645.
175. *Jager F., Greefkes J.* Frena ein System zur Übertragung von Sprache bei hohen Rauschpegeln.— *Phillips' techn. Rundschau*, 1957/1958, N 3.
176. *Jakobson R., Fant G., Halle M.* Preliminaries to speech analysis.— *MIT Acoust. Lab., Techn. Rept 13*, 1952.
177. *Jakobson R., Halle M.* Fundamentals of language. s' Gravenhage, Mouton and Co., 1956.
178. *Kacprowski J., Mikiel W.* Simplified rules for parametric synthesis of nasal and stop consonants in C — V syllables by means of the «terminal-analog» speech synthesizer.— *Acustica*, 1965/1966, v. 16, N 6, 356—364.
179. *Kahn L. R.* The use of speech clipping in single-sideband communication systems.— *Proc. IRE*, 1957, v. 45, N 8, 1148—1149.
180. *King J. H. Jr., Tunis C. J.* Some experiments in spoken word recognition.— *IBM, J. Res. and Developm.*, 1966, v. 10, N 1, 65—79.
181. *Klass P. J.* Computer responds to voice commands.— *Aviat. Week and Space Technol.*, 1961, v. 74, N 1.
182. *Kusch H.* Automatische Erkennung gesprochener Zahlen (Ziffern).— *NTZ*, 1965, Bd. 18, N 2, 57—62.
183. *Kusch H.* Das Segment ein Baustein der Sprache.— *NTZ*, 1967, Bd. 20, N 9, 495—501.
184. *Lehiste I., Peterson G. E.* Transition, glides and diphthongs.— *JASA*, 1961, v. 33, N 3, 268—277.
185. *Liberman A. M.* Some results of research of speech perception.— *JASA*, 1957, v. 29, N 1, 117—123.
186. *Liberman A. M., Ingemann F., Lisker L., Delattre P., Cooper F. S.* Minimal rules for synthesizing speech.— *JASA*, 1959, v. 31, N 11, 1490—1499.
187. *Licklider J. C., Pollack I.* Effects of differentiation integration and infinite peak clipping upon the intelligibility of speech.— *JASA*, 1948, v. 20, N 4, 42—51.

188. *Lindsey R.* Voice-controlled computer is «teachable».— *Aerospace Technol.*, 1967, v. 21, N 8, 52—53.
189. *Marcou P., Daguet J.* New methods of speech transmission. Information theory. London, Butterworths Sci. Publ., 1956, 231—244.
190. *Martony Y.* Some experiments on perceptual cues for swedish fricatives.— *Speech Commun. Seminar*. Stockholm, 1962.
191. *Mayer H. F., Holzler E.* Verfahren und Einrichtung zur elektrischen Nachrichtenübertragung.— *Pat. DBR N 878381*, 1953.
192. *Menon K. M. N., Jensen P. J., Donald D.* Acoustic properties of certain VCC utterances.— *JASA*, 1969, v. 46, pt 2, N 2, 449—457.
193. *Musmann H. G., Steiner K. H.* Phonetische Addiermaschine.— *Arch. elektr. Übertrag.*, 1965, Bd. 19, N 9, 502—510.
194. *Nagata K., Kato T., Shiba S.* Spoken digit recognizer for Japanese language, 4-th Internat. Congr. Acoust. Copenhagen, 1962. Congr. Rept, v. 1, s. a. NG 21.
195. *Nakata K.* Automatic recognition of speech.— *Elektronics and Commun. Japan*, 1963, v. 46, N 11.
196. *Nelson A. L., Herscher M. B., Martin T. B., Zadell H. J., Falter J. W.* Acoustic recognition by analog feature-abstraction techniques. Models for the perception of speech and visual form.— *Proc. Sympos. November 1964*, MIT Press. Cambridge, Mass., 1967, 428—440.
197. *Olson H. F., Belar H.* Phonetic typewriter.— *JASA*, 1956, v. 28, N 6, 1072—1081.
198. *Olson H. F., Belar H.* Phonetic typewriter III.— *JASA*, 1961, v. 33, N 11, 1610—1615.
199. *Olson H. F., Belar H., Rogers E. S.* Speech processing techniques and applications (synopsis).— *IEEE Internat. Convent. Rec.*, 1967, v. 15, N 1, 174—175.
200. *Öhman S. E. G.* Coarticulation in VCV utterances: spectrographic measurements.— *JASA*, 1966, v. 39, N 1, 151—168.
201. *Peterson E.* Frequency detection and speech formants.— *JASA*, 1951, v. 23, N 6, 668—674.
202. *Peterson G. E., Barney H. L.* Control methods used in a study of the vowels.— *JASA*, 1952, v. 24, N 2, 175—184.
203. *Petrick S. R., Willet H. M.* A method of voice communication with a digital computer.— *Proc. East Joint Comp. Conf. N. Y.*, 1960.
204. *Pichanick F. M. J.* Some aspects of the stop consonants.— *Trans. South African IEE*, 1960, v. 51, N 11.
205. *Potter R. K., Kopp G. A., Green H. C.* Visible speech Van Nostrand. N. Y., 1947.
206. *Potter R. K., Steinberg J. C.* Toward the specification of speech.— *JASA*, 1950, v. 22, N 6, 807—820.
207. *Reddy D. R.* Segmentation of speech sounds.— *JASA*, 1966, v. 40, N 2, 307—312.
208. *Sakai T., Doshita S.* The automatic speech recognition system for conversational sound.— *IEEE Trans. Electronic Computers*, 1963, v. 12, N 6, 835—846.
209. *Sakai T., Inoue S.* An analyzing equipment for the zero-crossing interval and its application to speech analysis.— *J. Inst. Electr. Commun. Engrs Japan*, 1956, v. 39, N 4, 404—450.
210. *Sakai T., Inoue S.* New instruments and methods for speech analysis.— *JASA*, 1960, v. 32, N 4, 441—450.
211. *Schneider H.* Die Verständlichkeit amplitudenbegrenzter Sprache.— *Frequenz*, 1956, Bd. 10, N 4, 97—106; N 5, 152—161.
212. *Schwartz M. F.* Transitions in American English (S) as cues to the identity of adjacent stop consonants.— *JASA*, 1967, v. 42, N 4, 897—899.
213. *Shatz C. D.* The role of context in the perception of stops.— *Language*, 1954, v. 30, N 1, 47—56.
214. *Sholtz P. N., Bakis R.* Spoken digit recognition using vowel-consonant segmentation.— *JASA*, 1962, v. 34, N 1, 1—5.

215. *Smith C. P.* A phoneme detector.— *JASA*, 1951, v. 23, N 4, 446—451.
216. *Smith J. E. K., Klem L.* Vowel recognition using a multiple discriminant function.— *JASA*, 1961, v. 33, N 3, 358.
217. *Steinbuch K.* Automatische Spracherkennung.— *NTZ*, 1958, Bd. 11, N 9, 446—454.
218. *Stevens K. N.* Towards a model for speech recognition.— *JASA*, 1960, v. 32, N 1, 47—55.
219. *Stevens K. N., House A. S.* Studies of formant transitions using a vocal tract analog.— *JASA*, 1956, v. 28, N 4, 578—585.
220. *Stevens K. N., House A. S., Paul A. P.* Acoustic description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation.— *JASA*, 1966, v. 40, N 1, 123—132.
221. *Suzuki J., Nakata K.* Phonemic classification and recognition of Japanese monosyllables.— *Electronics and commun. Japan (Denki Tsushin Gakkai Zasshi)*, 1963, v. 46, N 11.
222. *Suzuki J., Nakata K.* Recognition of spoken digits.— *Rev. Radio Res. Labs*, 1961, v. 7, N 31/32, 416—424.
223. *Suzuki H., Oizumi J.* The programme for the Japanese vowel recognition with learning capacity.— *Sci. Repts. Res. Inst. Tohoku Univ. B.*, 1963, v. 15, N 2, 7—27.
224. *Suzuki H., Oizumi J.* Statistical methods for learning of speech recognition.— *Internat. Conf. Microwaves, Circuit Theory and Inform. Theory*, Tokyo, 1964, pt 3, 139—140.
225. *Tarnóczy T.* Die akustische Struktur der stimmlosen Engelaute.— *Acta linguistica Hung.*, 1954, v. 4, N 3—4, 313—349.
226. *Terhardt E.* Beitrag zur automatischen Erkennung gesprochener Ziffern.— *Kybernetik*, 1966, Bd. 3, N 3, 136—143.
227. Voice-controlled manoeuvring for astronauts.— *JASA*, 1968, v. 43, N 6, 1468.
228. *Welch P., Wimpers R. S.* Two multivariate statistical computer programs and their application to the vowel recognition.— *JASA*, 1961, v. 33, N 4, 426—434.
229. *Wiren J., Stubbs H.* Electronic binary selection system for phoneme classification.— *JASA*, 1956, v. 28, N 6, 1082—1091.
230. *Yoshimura T.* Japan firm builds spoken voice digit recognizer.— *Electron News*, 1963, v. 8, N 390.

ОГЛАВЛЕНИЕ

Предисловие	5
Глава 1. Основные положения проблемы опознавания речевых сигналов	7
§ 1. Современное состояние проблемы	7
§ 2. Влияние характера произношения, индивидуальных особенностей и контекста на речевой сигнал	9
§ 3. О некоторых закономерностях восприятия речи человеком	12
§ 4. Выбор и характеристика направления исследований	16
Глава 2. Исследования клиппированного речевого сигнала	19
§ 1. Об использовании клиппированного сигнала при опознавании речи	19
§ 2. Влияние клиппирования речевого сигнала на разборчивость фонем	20
§ 3. О соотношении спектров речевого сигнала и его мгновенной частоты	28
Глава 3. Опознавание речи на уровнях сегментов и звуков	34
§ 1. Признаки речевого сигнала на уровне сегментов	34
§ 2. Вопросы сегментации речевого сигнала	39
§ 3. Опознавание гласных по квазистационарным сегментам	41
§ 4. Исследование переходных сегментов гласных. Признаки места образования и смягчения соседних согласных	52
§ 5. Опознавание щелевых и аффрикат по квазистационарным сегментам	67
§ 6. Опознавание взрывных по признакам шумовых сегментов	77
§ 7. К опознаванию фонем по последовательности сегментов	83
Глава 4. Опознавание ограниченного набора слов	89
§ 1. О признаках речевого сигнала на уровне слов	89
§ 2. Основные направления исследований	91
§ 3. Опознавание нескольких слов по последовательностям сегментов	96
§ 4. Опознавание нескольких десятков слов	111
Глава 5. Опознавание типовых фраз	122
§ 1. Об одном методе использования лингвистической информации	122
§ 2. Опознавание фраз при речевом управлении вычислительной машиной	125
Заключение	133
Приложение	135
Литература	138

Цемель Гири Израилевич

Опознавание речевых сигналов

Утверждено к печати Институтом проблем передачи информации

Редактор издательства Гальперин М. М. ■ Художник А. Д. Смеляков ■ Технический редактор И. Н. Жмуркина ■ Сдано в набор 7/І 1971 г. Подп. к печ. 24V/1971 г. Формат 60×90^{1/8}. Бумага № 2. Усл. печ. л. 9,5 Уч.-изд. л. 8,9. Тираж 3600 экз. Т-07489. Тип. зак. № 1864. Цена 53 коп. ■ Издательство «Наука». Москва К-62, Подсосенский пер., 21 ■ 2-я типография издательства «Наука». Москва Г-99, Шубинский пер., 10

53 коп.



ИЗДАТЕЛЬСТВО · НАУКА ·